Where are Emotions in Text? A Human-based and Computational Investigation of Emotion Recognition and Generation

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Enrica Troiano

aus Mutignano, Italien

Hauptberichter:	Prof. Dr. Roman Klinger
1. Mitberichter:	Prof. Dr. Sebastian Padó
2. Mitberichter:	Prof. Dr. Malvina Nissim

Tag der mündlichen Prüfung: 16 February 2023

Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart

2023

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Stuttgart, September 17, 2023

Place, Date

Enrica Troiano

Contents

List of Figures

List of Tables

Abstract

Since its early conception, artificial intelligence has strived to build machines that reason like humans and that approximate, among other things, our emotional intelligence. To sound as natural as possible, systems interacting with users must interpret, simulate and stimulate emotions, mastering an interpersonal competence that people apply in multiple communication channels. One of them is language. In verbal exchanges, internal affective states become observable objects that are amenable to the inquiry of natural language processing. My thesis develops within this computational framework and studies emotions expressed in written form.

Textual emotions have spurred much computational research, but their comprehension is not well-rounded yet. The literature on emotion recognition and generation mostly focuses on applicative objectives, with no solid tie to emotion theories. In consequence, works leverage conflicting premises about the type of data suitable to model, and how conspicuous its affective profile should be. Further, some approaches hinge on the connection between emotions and meaning, while others on their link to linguistic styles. This diversity of views suggests that much is still left to investigate at a fundamental level, to sharpen perspective on the expectations we set for machines.

This dissertation raises questions about two major theoretical gaps that hamper the creation of emotion-aware systems. Currently, the field lacks a clear understanding of (1) humans' abilities to detect emotions from text, and (2) the linguistic level that contributes to the emergence of emotions. Gaining insight on (1) how people infer emotions and (2) how these realize in text provides a comparison measure for the possibilities of the systems, as well as an idea of where they should model affective information. I address the two problems separately. Throughout the chapters, I establish trans-disciplinary connections, showing that a scientifically more inclusive account of emotions reflects their grounding in real-life events, and boosts their computational study.

The first part of the dissertation investigates the recognition performance of humans. I conduct multiple data creation activities that follow appraisal theories from psychology, which foreground the role of events in the emergence of emotions beyond language. To adapt this idea in the textual domain, I analyze judgments about descriptions of real-life circumstances (i.e., covert expressions with no emotion word like *"I received a promotion today"*). Further, I propose an annotation schema that allows to study both the agreement and the correctness of the coders. Emotion annotations turn out intelligible thanks to the underlying event evaluations, and their variability proves influenced by several extralinguistic factors. This group of studies illustrates that the subjectivity of emotions is not an obstacle for the creation of good quality data, but a fact that fosters our understanding of emotion mechanisms.

The second part of the thesis focuses on texts generated or annotated by systems. I use computational methods to examine where emotion phenomena locate in text, specifically to clarify their relationship with style and meaning, separately. First, I conduct a style transfer experiment based on backtranslation, aimed at investigating if the affective features of texts can be isolated from their meaning, as to give emotion a role of linguistic style. Second, I conduct a corpus-based analysis through event semantics. The study formalizes the link between events and emotions by leveraging frames, which confirm to incorporate many of the emotion features spelled out by appraisal theories in psychology. Overall, results support that the emotion phenomenon is polarized towards meaning, as long as this is considered a linguistic dimension separate from style, and it accounts for lexical relations and extralinguistic knowledge.

Zusammenfassung

Seit ihren Anfängen ist die künstliche Intelligenz bestrebt, Maschinen zu bauen, die wie Menschen denken und sich unter anderem unserer emotionalen Intelligenz annähern. Um so natürlich wie möglich zu wirken, müssen Systeme, die mit Nutzern interagieren, Emotionen interpretieren, simulieren und stimulieren und dabei eine zwischenmenschliche Kompetenz beherrschen, die Menschen in verschiedenen Kommunikationskanälen anwenden. Einer dieser Kanäle ist die Sprache. Im verbalen Austausch werden interne affektive Zustände zu beobachtbaren Objekten, die sich der Untersuchung im Bereich des Natural Language Processing unterziehen lassen. Meine Dissertation bewegt sich innerhalb dieses computergestützten Rahmens und untersucht Emotionen, die in schriftlicher Form ausgedrückt werden.

Textuelle Emotionen haben die Computerforschung stark beflügelt, aber ihr Verständnis ist noch nicht sehr ausgereift. Die Literatur zur Erkennung und Erzeugung von Emotionen konzentriert sich meist auf anwendungsbezogene Ziele, ohne eine solide Verbindung zu Emotionstheorien. Infolgedessen gehen die Arbeiten von widersprüchlichen Prämissen aus, was die Art der zu modellierenden Daten und die Auffälligkeit ihres affektiven Profils angeht. Darüber hinaus konzentrieren sich einige Ansätze auf die Verbindung zwischen Emotionen und Bedeutung, während andere ihre Verbindung zu linguistischen Stilen betonen. Diese Vielfalt der Ansichten bedeutet, dass es noch viel zu erforschen gibt, um den Blick für die Erwartungen zu schärfen, die wir an Maschinen stellen.

Diese Dissertation wirft Fragen zu zwei großen theoretischen Lücken auf, die die Entwicklung emotionsbewusster Systeme behindern. Gegenwärtig fehlt ein klares Verständnis (1) der menschlichen Fähigkeit, Emotionen aus Texten zu erkennen, und (2) der linguistischen Ebene, die zur Entstehung von Emotionen beiträgt. Ein Einblick in (1) die Art und Weise, wie Menschen Emotionen erkennen und (2) wie sich diese in Texten realisieren, bietet einen Vergleichsmaßstab für die Möglichkeiten der Systeme sowie eine Vorstellung davon, wo sie affektive Informationen modellieren sollten. Ich gehe auf diese beiden Probleme getrennt ein. In den einzelnen Kapiteln stelle ich transdisziplinäre Verbindungen her und zeige, dass eine wissenschaftlich umfassendere Darstellung von Emotionen ihre Verankerung in realen Ereignissen widerspiegelt und ihre computergestützte Untersuchung fördert.

Im ersten Teil der Dissertation wird die Fähigkeit von Menschen untersucht, Emotionen zu erkennen. Dafür erstelle ich mehrere Datensätze. Die Datensets basieren dabei auf *appraisal* Theorien aus der Psychologie, deren Fokus es ist die Rolle von Ereignissen bei der Entstehung von Emotionen jenseits der Sprache in den Vordergrund zu stellen. Um diese Idee auf den textuellen Bereich zu übertragen, analysiere ich Beurteilungen von Beschreibungen realer Lebensumstände (d. h. verdeckte Ausdrücke ohne Emotionsworte wie "*Ich wurde heute befördert*"). Außerdem schlage ich ein Annotationsschema vor, mit dem sowohl die Übereinstimmung als auch die Korrektheit der Annotator:innen untersucht werden können. Die Emotionsannotationen werden dank der zugrunde liegenden Evalutation der Ereignisse verständlich und ihre Variabilität wird durch verschiedene außersprachliche Faktoren beeinflusst. Diese Gruppe von Studien veranschaulicht, dass die Subjektivität von Emotionen kein Hindernis für die Erstellung von Daten guter Qualität ist, sondern ein Aspekt, der unser Verständnis der Emotionsmechanismen fördert.

Der zweite Teil der Arbeit konzentriert sich auf Texte, die von Systemen generiert oder annotiert werden. Ich setze computergestützte Methoden ein, um zu untersuchen, wo Emotionsphänomene im Text verortet sind, insbesondere um ihre Beziehung zu Stil und Bedeutung getrennt zu klären. Zunächst experimentiere ich mit Methoden des Style Transfer und Rückübersetzungen. Hier untersuche ich, ob die affektiven Merkmale von Texten von ihrer Bedeutung isoliert werden können, so dass Emotionen eine Rolle im sprachlichen Stil spielen. Zweitens führe ich eine korpusbasierte Analyse mittels Ereignissemantik durch. Die Studie verbindet Ereignissen und Emotionen durch den Einsatz von Frames. Hierbei zeige ich, dass Frames viele der Emotionsmerkmale enthalten, die in den *appraisal* Theorien der Psychologie beschrieben werden. Insgesamt belegen die Ergebnisse, dass das Phänomen der Emotionen eng mit der Bedeutung verbunden ist, sofern diese als eine vom Stil getrennte sprachliche Dimension betrachtet wird und lexikalische Beziehungen und außersprachliches Wissen berücksichtigt werden.

Acknowledgements

If this manuscript has come to light, it is thanks the support of my supervisors, Roman Klinger and Sebastian Padó. They guided me through research with the finest taste for detail and an eye on my broader path. Having them as mentors is an opportunity that I have never taken for granted.

I am also grateful to Malvina Nissim, for reviewing this thesis and asking thrilling questions during my defense; Sabine Schulte im Walde, for her wise pieces of advice; and Kai Sassenberg, whose ideas have significantly enriched my interdisciplinary studies.

A great deal of my work owes to the input of Ada Lorenzo, for she has shaped how I think about emotions, and to the direct or indirect contribution of *each* member of IMS. I thank all the students who have crossed my PhD path, those I have collaborated with, Martin Rettig for his technical assistance, as well as Diego Frassinelli, Laura Oberländer and Agnieszka Faleńska, who have taught me that research is much more joyful when it is done with friends.

Lastly, I wish to express immense appreciation to Gabriella Lapesa, who has sided me with constant and graceful care.

Chapter 1

Introduction

Ways of thinking (Minsky, 2006), gut reactions (Prinz, 2004), judgments of values (Nussbaum, 2004): these are few of the ways in which emotions, pervasive and yet not fully understood "things" that humans feel, have been referred to across research fields. Today, competing perspectives on emotions come from diverse disciplines, such as psychology, which attempt to explain the involvement of the brain and body in the subjective core of our experiences. Emotions are large networks of processes putting the external and internal worlds into contact, affections that fill in matters of facts with meaning, that we communicate to others, and whose expression can change how others in turn will think or behave (Buechner et al., 2015; Zhao et al., 2020; Van Kleef et al., 2011).

This link with communication emerges in various modalities. Emotion episodes can be given away by one's bodily gestures and facial expressions, but they become public all the same in the sphere of language (Fussell, 2002). We can describe the passing of an exam as a joyful academic turning point, and the fight with a friend in terms of an infuriating event. Ultimately, understanding what people feel from the words that they choose is key to successful interactions, as it serves to grasp the mental states of our interlocutors and, often, the very gist of their utterances (Scheff, 1973).

If words reflect our experience of the world, any semantic account of language has a good reason to investigate the emotions that accompany such experiences. Indeed, around twenty years ago, verbal emotions evolved into an area of inquiry for Natural Language Processing (NLP), a subfield of artificial intelligence. Computational emotion analysis has aimed at creating machines that "recognize, express, model, communicate, and respond to emotional information" (Picard, 2000), like other areas focused on affective computing, and has done so using linguistic data.

The computational study of emotions in language has progressed enormously since then (e.g., Felbo et al., 2017; Ghosh et al., 2017; Majumder et al., 2020), but its goal to devise a fully-fledged emotion machine that understands and produces texts with an affective signature remains only partially fulfilled. Many fundamental questions are still open for investigation. For instance, how do emotions transpire through language? Do they characterize any type of text? It is intuitive to think that words possess some sort of emotionality that reveals how facts affect us, but this seemingly commonsense knowledge deserves serious consideration. What is at stake is not only a theoretical answer, but a critical re-thinking of the data and approaches settled in the field (e.g., Are news headlines an appropriate source to model emotions?).

Therefore, this thesis tackles the theoretical basis in the study of linguistic affect. It pushes forward a discussion regarding some best practices to exploit human knowledge of emotions in text, highlighting important points to devise emotion-oriented computational models. More precisely, it covers three macro topics: how people recognize emotions in text, what plays a role in such a task and, finally, where emotions come to emerge in language. All of them intersect my underlying attempt to bring computational linguistics closer to psychology.

1 Psychological vs. NLP Perspectives on Emotions

Emotions stand out in the landscape of affect. For one thing, they are many. Joy, sadness, fear, and their relative gradable versions like serenity–ecstasy, pensiveness–grief, apprehension–terror, represent only a small part of virtually everyones' emotional repertoire, while a phenomenon like sentiment boils down to a handful of categories (e.g., neutral, negative, positive). Further, every emotion has a relational nature. We are happy, sad, scared, *of* the stimulus that causes us to feel so, *about* it or *because of* it. Lastly, and perhaps most importantly, it is difficult to pinpoint what emotions are. For this reason, the search for the emotion elementary particles has spurred an extremely diverse literature, from early mechanistic views in philosophy (Descartes, 1989) to contemporary approaches in neuroscience (Panksepp, 2004).

Psychology has nonetheless found a few points of consensus. One is the idea that emotions can be studied systematically (cf. Dixon, 2012, p. 338). The other is the observation that they are recognizable through at least some "diagnostic features" (Scarantino, 2016). These features are constant to all emotion episodes, always liable to be broken down into:

- the presence of a triggering **event**;
- an **evaluation** of the event that individuals conduct based on their goals, values, memories, morals and preferences;
- several **concomitant changes** which can be visible (e.g., crying) or not (e.g., the heartbeat goes faster).

Emotions thus derive from the assessment of the qualities of a stimulus, which prompt a corresponding qualitative state in their experiencer.

Investigating the interpretation and production of these states in text is the goal of the NLP subfields of computational emotion analysis and computational emotion generation. The first casts the task of automatic emotion recognition (Zhang et al., 2020a; Alvarez-Gonzalez et al., 2021; Guibon et al., 2021), the other is concerned with automatic affective writing (Huang et al., 2018; Goswamy et al., 2020). Most efforts have gone in the analysis direction (taken also in this dissertation). It consists in modeling the import of a text that (supposedly) corresponds to what humans feel outside the domain of language.

Emotion recognition, also called "emotion detection", assigns texts to categorical emotion labels (Mohammad, 2012; Klinger et al., 2018, i.a.) or to dimensional features (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017a, i.a.). Works exploit theoretical insights about which emotions should be considered (e.g., anger, sadness, fear) and how to describe them (e.g., with discrete labels). In this light, the emotion experiences that scientists focus on are the connective tissues between NLP and psychology, but the two fields differ more than they share similarities. Theoretical research looks for first principles to explain emotions within and between people; computational treatments of language refrain from providing a precise characterization of emotions, bypassing the difficulty to frame a clear-cut object of study. That appears, for instance, in the annotation activities that assign labels to textual data, as a fundamental step to learn models, as well as standalone research on emotions. These labels correspond to the impressions of humans about the emotions inferable from text. To gather them, researchers rely on the knowledge that laypeople have from their own experiences, and do not put into question how well it applies to language or to the particular textual domain of interest.

Certainly, it is not trivial to find a transparent definition amidst the patchwork of theoretical proposals, but taking a relaxed approach towards the "thing" being investigated muddles the picture as to whether it is the same across studies and textual varieties (e.g., How to compare the affective reaction elicited by a tweet to that of a poem?). Besides, psychology can lend more than labels to classify and affective dimensions to rate – it suffices to think of the dynamics that take place under the blanket of any emotion, with all contributing components and diagnostic features. An idea that has found little space in the computational agenda is the relationship between emotions and the evaluation of events. The ability to evaluate an environment allows humans to figure out its properties (if it is threatening, harmless, requires an action, etc.), which in turn determine if and how they react emotionally. To overlook evaluations is to dismiss a primary emotion resource, and above all, an account of emotions for what they are: grounded phenomena.

Summing up, there is a discrepancy between the study of emotions in vs. out of language, with computational linguistics preferring a pragmatic approach that remains on the surface of what has to be found in a text, and psychology expanding on what lies underneath its embodied occurrence.

2 Challenges for Computational Emotion Analysis in Text

Despite the simplifying assumptions they make about emotions, recognition-based approaches have not fully solved their task. At least three groups of challenges can be identified in the field, evidence that emotions still motivate the study of their linguistic realization, alongside the psychological inquiry of their embodied emergence.

Automatic Emotion Recognition Challenges. Automatic NLP systems should sense emotions like humans do, but this goal is severely demanding. One reason is that emotions can be conveyed covertly, without being directly named, via texts that contain no mental state nor evaluative attitude at all. The sentence *"I was grinning from ear to ear"* is illustrative of how joy can turn into an **implicit expression**, which can be conceptually separated from **explicit expressions** (e.g., *"I'm happy"*) containing words that signify an emotion or a personal involvement in some states of affairs. Emotion analysis assumes that affective meanings can be inferred from both types of texts. That is feasible for humans: people understand emotions from their interlocutors even when none is mentioned. However, such interpretations are drawn via pragmatic inference (Grice, 1975), presupposing a bundle of extralinguistic knowledge that systems do not necessarily possess.

The transition from explicit to implicit expressions is abrupt, because of the need to integrate knowledge with appropriate data for the systems. Strategies used by humans to infer emotions from implicitly emotional texts could inform modeling approaches. Hence, the first step towards that goal is to understand how well humans perform the task, and on which texts it can be performed.

Annotation Challenges. Much data in the field is extracted from already-available sources, either online platforms (Wood et al., 2018; Liew, 2014) or established textual corpora (Mohammad, 2011; Esuli et al., 2008). Having been produced for other purposes, the data often comes unlabeled for emotion classification. The texts stand in need of an association with the one or many emotions that they "contain", but for a researcher to gain contact with the texts' writers and ask about the correct emotion interpretation of their production is time-consuming, or simply unfeasible.

As a solution, annotation efforts are accomplished with the help of readers (e.g. Edmonds and Sedoc, 2021; Li et al., 2016a; Quan and Ren, 2009). These can return an extremely varied and inconsistent picture of the emotions present in the data, because emotions are based on personal evaluations contingent on the needs, personal values, desires and other criteria of importance for each, separate individual (the absence of an emotion definition discussed above intensifies the diversity of the collected judgments). Low agreement among annotators is commonly treated as an inherent flaw of judgments about emotion meanings. In this respect, they differ from judgments on semantic phenomena that leave less space for idiosyncratic and world-driven intuitions, like named entity recognition (Balasuriya et al., 2009, i.a.,) or anaphora resolution (Goecke et al., 2008, i.a.,).

This represents a concrete problem. For a corpus to become machine-learning useful, the annotations are typically aggregated into one or many labels (Bobicev and Sokolova, 2017; Štajner, 2021), and disagreements render the job difficult. Especially if no ground truth is provided by the writers, researchers must face a methodological decision to reconcile the crowd's understanding, and to adjudicate the labels that represent it the most. Therefore, a second challenge for computational emotion analysis is to clarify how to deal with the difficulty to obtain acceptable agreement scores.

Challenges to the Understanding of Linguistic Realizations. By loosely referring to textual emotions as "emotion meanings" above, I have taken a precise stance that is worthy of debate. The composition of the terms "emotion" and "meaning" suggests that the former is part of a language's semantics. This viewpoint is made explicit by dictionary-based approaches to emotion analysis. They advocate that certain words are endowed with a prototypical affective connotation, which would permit an immediate comprehension of, e.g., "die" as loaded with sadness, "win" with joy, "ghost" with fear, and so on. However, that emotions are part of meanings is an assumption.

There is also another potential equivalence to uphold in addition or in alternative to the emotion–meaning one. Emotions can be treated as the *style* of utterances, as suggested by the intuitive possibility to find texts that are semantically similar but different in their emotional import.

Relief: "*The fire that burned the wood was extinguished before it hit the houses.*"

Anger/Sadness: "*The houses are safe, but the forest is devastated.*"

The sentence (arguably) laden with relief is paraphrased into one with a more negative emotional connotation. This shows that the core information of texts can be repurposed into a new emotion.

The automatic generation of paraphrases that display an emotion different from the input has started to take hold in NLP, resonating with recent works of style transfer (Jin et al., 2022). The idea is that style is independent of the utterances' gist, and can therefore be modified at will to re-style meaning into multiple affective facades.

Emotion analysis has capitalized on the semantics- or style-centered viewpoint but has never tested either one. As a result, it is unclear if emotions are a dimension of meaning or of other sides of our verbal productions; this translates into a practical confusion in respect to the linguistic level at which they can be investigated. Is emotion style transfer doable? Does that overshadow computational approaches that study emotion meanings?

3 Thesis Contribution

Overall, a lack of clarity dominates different steps in the workflow of research in emotion analysis, starting from the compilation of the data of interest, specifically regarding what to collect and how to annotate it, and its automatic processing, with the issue of how such data can legitimately be used. The field is need of a more robust understanding of the object it investigates in text. My thesis addresses this problem.

3.1 Research Questions

My contributions develop around three research questions at varying degrees of abstractness in a linguistic perspective. I ask:

RQ1: How well do humans recognize emotions from implicit expressions? I observe the extent to which readers infer the correct emotions, by focusing on factual statements (more precisely on event descriptions) in the role of well-defined implicit expressions. The link between emotions and events is documented outside the verbal domain. Knowing that, I investigate their link in text. The goal is to understand if implicit expressions are suitable data to investigate the emotion recognition ability of humans, and consequently to learn automatic models.

RQ2: How do in-text and beyond-text factors affect human emotion recognition? Which is to say: how to make the most out of the subjectivity of emotion judgments? To deal with the diversity of emotion inferences from text, I study aspects that pertain either to language or to its users. The aim of this step is the adoption of additional information besides emotion judgments, such as people's demographics and current emotion states, as useful means to explain disagreements.

RQ3: Where are emotions? Understanding the linguistic level at which emotions realize can give fruitful insight into the information relevant to grasp emotions in text. Therefore, this question contributes to the theoretical knowledge about how emotions, which are prominently cognitive phenomena, turn into linguistic phenomena; from a more applicative standpoint, it determines whether modeling emotions in terms of style or of word meanings are feasible computational tasks.

3.2 Approach and Answers

The thesis positions itself within NLP but looks for points of convergence with psychology and general linguistics, and exploits tools from the three of them. From NLP, I utilize data analysis strategies and computational methods that classify and generate text; I also make use of crowdsourcing and in-lab practices for data labeling, conducted both by humans and with automatic annotators to process emotions at different levels of granularity (i.e., as a rich set of categories and as *emotionality* – whether a text has an emotion, irrespective of what that is). From psychology, I tap on some theoretical models which delineate the mechanisms underlying emotions beyond language, and I adapt them to the linguistic domain. Further, I touch upon linguistics interests, specifically regarding the (apparently) dichotomous notions of style and meaning.

Approach to answer RQ1. The discussion starts from the premise that emotions are *emotions of embodied agents affected* by events. To formalize this idea, I sketch a communication framework that builds on top of the involvement of two actors in the transmission of emotion signals (one who produces the message, and another interpreting it). I further bridge it with a class of psychological models that characterize emotions around events, which allow analyzing the behind-the-scene of an emotion judgment (i.e., event evaluations). In practical terms, I use this framework to collect the largest corpus of event descriptions annotated with emotions, a number of personal factors (e.g., personality traits, familiarity with the texts' topics), and event evaluations, from the perspective of the text writers and readers. Moreover, I compile the first multilingual corpus of event descriptions associated with emotions, as provided both by first-hand experiencers and by external readers.

I find that readers recognize emotions imperfectly. That is expectable, given the subjectivity of the task. However, I show that it is possible to at least assess the quality of judgments about implicit expressions with an experimental design based on the chosen communication framework. I further highlight that event evaluations boost a researcher's understanding of the collected emotion annotations, because they can stand as a *justification* of emotion judgments.

Approach to answer RQ2. Also the answer to the second research question leverages humans' annotations. I address the subjectivity of emotions to investigate the annotation quality of readers. I illustrate that more annotation layers than emotion- and event evaluation-centered ones give insight into the judges' performance. Such layers regard personal and textual factors based on psychological theories and findings.

Some disagreements turn out to be non-random. In fact, they relate to the considered factors, calling for better measures of inter-annotator agreement and for the inclusion of multidimensional annotations (e.g., the demographics of the coders, their own emotion state) that fit the complex nature of verbal emotions.

Approach to answer RQ3. I describe two separate experiments dedicated to different levels of linguistic realization: style and meaning. First, I verify if the claim that emotions are a linguistic style stands up to the test of style transfer, namely, the possibility of modifying a text's style while leaving its meaning untouched. This emotion-aware paraphrasing task is addressed with the help of machine translation. I obtain quantitative evidence that an ordinary machine translation transformation loses emotions, and for this reason, backtranslations can be used for style transfer. On the other hand, with a qualitative analysis of the generated texts, I highlight that machines confirm to be a successful transfer what hardly passes human inspection, casting doubt on the emotions-style identity (when style is deemed a dimension orthogonal to meaning).

Next, I embrace the stance that emotions belong to the meaning of utterances, close in spirit to dictionary-based studies. With an experiment that makes use of computational methods for the annotation of an unlabeled corpus, I show that emotions are part of semantics, but not in the lexically-focused sense of dictionary-based approaches. They are so in a U-semantic perspective, i.e., one that accounts for the relation between the linguistic units in a text, their context and the process of their interpretation (Fillmore, 1985). As such, they are captured by frames, which are formal ways to represent a framework of meaning (e.g., "beat", "defeat", "demolish", "prevail" are related concepts forming a coherent structure in the frame BEAT_OPPONENT, and the "change", "swap", "switch", "trade" of an artefact are all instantiations of the conceptual suit-case of the frame CHANGE_TOOL). The concrete output of this study is a dictionary of associations between frames and emotionality (e.g., the frame BEAT_OPPONENT is packed with emotion, the frame CHANGE_TOOL is not). In addition, I elaborate on a detailed qualitative analysis. Frames that are intrinsically emotional prove to pick up the several ways in which a text can express affectivity, and corroborate the communication framework instantiated for RQ1. Among other things, they mirror people's event evaluations (considered in RQ1 as well), confirming once more that psychological theories with an event-centered view of emotions are advantageous for exploring affect in language.

Overall, by answering RQ1 and RQ2, I sustain the idea that theoretical considerations from psychology improve computational emotion analysis; with the last question, I reverse the direction of inquiry.

4 Thesis Plan and Publications

On a high level, Chapter 3 and Chapter 4 belong to the same narrative arc about working with humans for the computational purpose of emotion analysis; Chapter 5 and Chapter 6 can be grouped side-by-side in the discourse on "where" emotions emerge in text.

Much material in this thesis appeared in publications that I contributed to during my doctoral studies. Since the articles are the result of joint work, they will be presented in terms of *our* studies and of findings that *we* found. Each chapter either reports on one article or is a synthesis of many, as follows.

Chapter 2 provides preliminary background. I present the psychological models of emotions that can be found in computational emotion analysis, and I linger on *appraisal theories*, which are the psychological centrepiece of my work. I introduce corpora as well as practices for emotion annotation and the consequent computation of inter-annotator agreement. Further, I overview the task of style transfer, with a special focus on the objectives that concern the (supposed) style of emotion, and I describe the semantics of frames.

The portion of the chapter dedicated to style transfer is based on:

- Troiano, E., Velutharambath, A., and Klinger, R. (2022b). From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*, pages 1–60

  There, I actively took part in the organization of styles into a hierarchy that encompasses the style transfer literature, I summarized and linked technical and theoretical works, and provided a critical discussion of the task limitations.

Chapter 3 addresses RQ1. As the starting point of my contribution, the sense in which texts have an *affective* side will be put into focus, to understand what we talk about when dealing with emotions in verbal communication. I define a data collection procedure instantiated with crowdsourcing. By leveraging the resulting corpora, I emphasize the limits of emotions recognition from implicit expressions, the advantage of making use of this specific crowdsourcing procedure, as well as the benefit of collecting annotations concerning event evaluations.

The chapter combines two publications:

- Troiano, E., Oberländer, L., and Klinger, R. (2023b). Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72

  For this work I collected and curated the data.

- Troiano, E., Padó, S., and Klinger, R. (2019). Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics

  In this work, I defined the crowdsourcing guidelines, carried out the data collection step, analysed the annotations and implemented the systems for emotion classification.

The discussion also builds on top of results that I do not elaborate on in detail, published in the following articles:

- Troiano, E., Oberländer, L. A. M., Wegge, M., and Klinger, R. (2022a). x-enVENT: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association

I took care of the design of the study, guidelines definition, corpus construction, and analysis.

- Hofmann, J., Troiano, E., and Klinger, R. (2021). Emotion-aware, emotion-agnostic, or automatic: Corpus creation strategies to obtain cognitive event appraisal annotations. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 160–170, Online. Association for Computational Linguistics

  My contribution was both conceptual (in the theoretical discussion to link appraisal dimensions to the emotion analysis goals) and practical (I participated in the outline of guidelines and in data annotation).

Chapter 4 answers RQ2. After identifying a handful of candidate aspects important for emotion recognition, I compare the way they reflect on inter-annotator agreement, and expand on how the (learned) entanglement with disagreements can be handled by future studies.

- Troiano, E., Padó, S., and Klinger, R. (2021). Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics

  I was responsible for guideline definition, annotation collection, data analysis, and modeling.

Part of this chapter also refers back to:

- Troiano, E., Oberländer, L., and Klinger, R. (2023b). Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72

With the final two chapters, I move my attention to RQ3. **Chapter 5** presents the results of emotion style transfer following:

- Troiano, E., Klinger, R., and Padó, S. (2020). Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, Barcelona, Spain (Online). International Committee on Computational Linguistics

I implemented the experimental setup and conducted all analyses.

Chapter 6 turns to the last experiment. It concerns the relationship between emotionality and frames based on:

- Troiano, E., Klinger, R., and Padó, S. (2023a). On the relationship between frames and emotionality in text. *Northern European Journal of Language Technology*, 9(1)

  The practical implementation of the experiments, the qualitative and quantitative analysis were my contribution.

Chapter 7 concludes this dissertation, with a summary of the filled-in research gaps, the findings, the limitations of my experiments, and the questions for future research that they elicit.

The thesis is also the product of work I do not report here, but which helped me navigate the problems of the various chapters. They are:

- Hofmann, J., Troiano, E., Sassenberg, K., and Klinger, R. (2020). Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics

  I contributed to data annotation and study design.

- Helbig, D., Troiano, E., and Klinger, R. (2020). Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 41–50, Online. Association for Computational Linguistics

  My involvement consisted in supervising the execution of the task, cooperating in the human evaluation of the style transfer outputs, and the final qualitative analysis.

- Sabbatino, V., Troiano, E., Schweitzer, A., and Klinger, R. (2022). "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 37–50, Dublin, Ireland. Association for Computational Linguistics

  My contribution relates to the conceptual formulation of the task.

- Barnes, J., Oberlaender, L. A. M., Troiano, E., Kutuzov, A., Buchmann, J., Agerri, R., Øvrelid, L., and Velldal, E. (2022). SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics

  My role as a co-organizer of the shared task was to analyze qualitative patterns across the outputs of the submitted systems.

Chapter 2

Background and Related Work

Most works in computational emotion analysis start from the selection of a theory of reference from psychology, which determines the emotions to study and the ways to represent them. In turn, theories originate from different premises about the important aspects of emotions. To establish the major themes underpinning the dissertation, I provide an overview of the psychological lines of research that have affected emotion analysis (Section 1), highlighting their pitfalls and advantages. I then link them to the corresponding computational approaches in text (Section 2) which have created corpora or dictionaries for the task of emotion recognition.

Corpora have also supported style transfer, the generation task to which Section 3 shifts attention. There, I show how emotions fit a hierarchy that organizes the styles found in the literature, and I review the existing publications on emotion style transfer. Dictionaries, on the other hand, bring up a discussion of lexical meaning. Section 4 introduces frame semantics as a formalism to represent events, and for that reason, particularly suitable to study emotions.

Highlights

Events and evaluations are key to the definition of emotions and their study in language.

Emotions have been a style for the task of style transfer.

Emotion recognition beyond text is influenced by multiple factors.

Frame semantics is a powerful framework to describe events.

1 Emotions in Psychology

Emotions are an interdisciplinary matter. Since they echo in our brain, body, and actions, they promote the study of the human condition in various regards, from the neuroanatomical substrates of behavior to the principles of humans' development and capabilities (Kavanaugh et al., 1996). This subject has somewhat of a primary connection to psychology, among many disciplines. Emotions put into motion the mind–body system in its entirety, involving cognitive, physical, social abilities, to name just a few, which are the motif of research par excellence in the field.

One of the things psychology has investigated is how emotions arise. Its academic production elaborates profusely on their eliciting conditions, their component mechanisms, and the stereotypical expressions that develop in their company (Scherer, 2000; Gendron and Feldman Barrett, 2009). Actually, the field has considered more than such a private, within-individual dimension of emotions. Giving an all-round account of these mental facts, it has had a say about them being public objects experienced by observers: as a person feels an emotion, another might recognize it.

The current section addresses past work conducted in psychology on these two themes: the emergence of emotions and their recognition.

1.1 Three Traditions

Research explaining how emotions arise accounts for different types of emotion episodes, motivated by specific assumptions on the emotion causes (Cannon, 1927) and effects (Tomkins, 1962), their evolutionary relevance and universal validity (Ekman, 1992; Plutchik, 2001), as well as their underlying dimensions (Russell and Mehrabian, 1977). Although disagreeing, all such perspectives can be sorted into three large streams of thought, inside the traditions that equate emotions with *motivations*, *feelings*, and *evaluations* respectively (Scarantino, 2016). For the first two, I will only introduce the gist, while I will dissect the details of the third, which will recur throughout the thesis.

1.1.1 Motivational

A great deal of theories stems from the evolutionary significance of emotions (Izard, 1971; Tooby and Cosmides, 2008). That can be understood by looking at their link to behavior. Even when emotions do not prompt any concrete action, they still cause changes in action readiness (Frijda, 1996). Emotions are thus *motivational states* because they predispose a response to reach the objective one cares about, modifying this way one's relationship with the environment (Dewey, 1894). They move us, they are a drive that pushes us to accomplish fundamental everyday needs and "that, in the course of our evolution, has done better than other solutions in recurring circumstances that are relevant to our goals" (Ekman and Cordaro, 2011, p. 364).

Emotions help, for instance, to communicate socially relevant information, for they trigger specific physiological symptoms (e.g., smiling, frowning) and patterns of behavior. This correspondence between bodily manifestations and emotions pushed research in psychology to group emotional states into a few natural language categories (smiling can indicate *joy*, frowning can signal *anger*, etc.). The motivational tradition embraces precisely this idea. It encompasses theories proposing inventories of emotion names, which correspond to some primary (i.e., biologically basic) motivational states.

Scientists think in various ways about the characteristics of basic emotions, and in consequence, they end up analyzing different sets thereof. One well-established exemplar is Ekman's (1992), an inventory that lists six categories: anger, disgust, fear, joy, sadness, and surprise. In the Darwinistic approach of Ekman, basic emotions are observable and measurable phenomena. They all show a quick onset, a brief duration, an unbidden occurrence, and distinctive thoughts (Ekman, 1999). Most importantly, they are innate and universal. Ekman notices their occurrence among other primates and across cultures, where an emo-

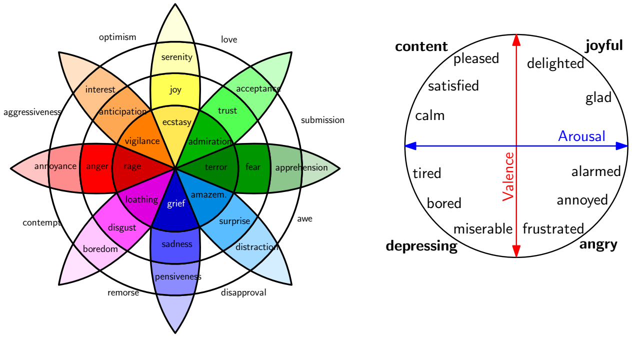


Figure 2.1: Left: Plutchik's Wheel of Emotions, a basic emotion model. Right: The circumplex model of emotions, a dimensional emotion model.

tion program always initiates the same changes. Facial expressions are a case in point. Based on the assumption that they are an inescapable effect of emotions, many studies have coded the mapping between specific configurations of facial muscles and specific emotion states (Clark et al., 2020) – e.g., cheek raising, outer lip corners pulling, teeth showing are physiological signatures of joy (Du et al., 2014).

Another motivational model that extends Darwin's ideas is the Wheel of Emotions of Plutchik (2001), shown in Figure 2.1 (left). Plutchik identifies fundamental or primary emotions on the basis of classes of evidence about their evolutionary fitness. He notices that the requirements imposed by the environment upon living organisms are always carried out with a few patterns of behavior. These prototypes, as Plutchik calls them, (1) are adaptive patterns linked to survival with generality across all phylogenetic levels, (2) serve a function (e.g., to protect, to reject, to destruct) that re-establishes a state of rest for the organism and reduces disequilibrium in a state of emergency, (3) have a bipolar structure (e.g., protection is antipodal with rejection), and (4) can be coupled with emotion words in language (Plutchik, 2001, 1984).

Eight primary emotions comply with these conditions. Each of them matches a behavior of either protection (fear), destruction (anger), reproduction (joy), deprivation (sadness), incorporation (acceptance), rejection (disgust), exploration (anticipation) or orientation (surprise) (Plutchik, 1970). The wheel arranges them into sets of opposite pairs, as diametric ends (i.e., anger vs. fear, joy vs. sadness, surprise vs.

anticipation, trust vs. disgust). Further, it illustrates how different fundamental emotions can occur together, merging like colors to form others (e.g., a composition of trust and joy results in love).

An important feature of Plutchik's emotions is their intensity gradation, which varies along the wheel's radius (e.g., higher intensity, darker color: ecstasy; lower intensity, fairer gradation: serenity). Intensity covers multiple factors of emotions, such as impulses to action, revisions of beliefs, the disregard/pondering of the emotion-triggering event, but in a nutshell, it can be thought of as the magnitude of the subjective feeling that characterizes an emotion (Frijda et al., 1992). Lacking in the framework of Ekman, the continuous variable of intensity bridges Plutchik's discrete emotion model with the *feeling tradition*, where emotions are defined in a space that possesses a dimension similar to intensity, i.e., the dimension of arousal.

1.1.2 Feeling

In its early version, this stream of research purported that "we feel sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful" (William James, reported from Myers, 1969). Pioneered by James, the argument that emotions have physiological causes puts again on the table a discussion about the emotion–body link, but under a radically different light than it has in the motivational account. Rather than effects, bodily signals assume a preliminary function for any emotion episode: objects and facts (not emotions) excite changes in the body, like reflexes; emotions are *feelings* of such changes (James, 1894).

Although controversial (Golightly, 1953), James' proposal outlines an idea that has stood the test of time: feeling is a mode of perception common to all emotions. Recent thinkers still profess it, but with some fundamental adjustments. First off, they have abandoned the perception-to-emotion causality (Averill, 1980; Oatley, 1993; Barrett and Russell, 2015), which ignores that not all emotional processes develop in that strict sequential order. In addition, they treat emotions as *constructs* shaped dynamically, where the body brings a relevant, yet partial, contribution. The prominent position there is held by *core affect*, a conscious sense of feeling that always accompanies people (Barrett, 2017) and which allows them to answer the question "how do you feel?" at any point in time (Russell, 2012). Experiencing an emotion means perceiving the fluctuations of the core affect system, which occur along a finite number of components. First “is how pleasant or unpleasant you feel, which scientists call valence. [...] The second feature of affect is how calm or agitated you feel, which is called arousal” (Barrett, 2017, p. 72). Being afraid, for instance, would result from the interpretation of a state of high arousal and displeasure that an experiencer finds herself in, and from categorizing such a condition under the concept of fear (Scarantino, 2016).

A conflation of some positions of the core affect space into emotion tokens is illustrated in Figure 2.1 (right). The image represents the circumplex model of valence[1] and arousal of Russell (1980), in which Posner et al. (2005) operate the continuous-to-discrete mapping. Compared to this model, Bradley and Lang (1994) add a dimension of dominance (i.e., the power that an experiencer self-perceives in a situation), defining a space where emotions vary from one another along the three VAD (Valence-Arousal-Dominance) components.

Affect-oriented approaches in the feeling tradition actually diverge on the categorization of the continuous affective space. Some of them regard it as not strictly necessary (Russell, 2012), because it only influences one’s realization of her emotional experience but not the experience itself. Overall, however, constructionists concur on a key tenet: the organism does not possess a single emotion module. On the contrary, it uses many resources that engage variously in different occasions, leading to the emergence of alternative emotions much in the way that the building blocks of an algorithm can be re-purposed to create alternative instructions (Feldman Barrett, 2017, i.a.).

This view resists the one-to-one correspondence between emotions and behavioral or anatomical reactions assumed by motivational theorists. Therefore, it sets the feeling tradition apart from the idea that “emotions have ontological status as causal entities [that] exist in the brain or body and [that] cause changes in sensory, perceptual, motor, and physiological outputs” (Barrett, 2005). In line with James’ theoretical system, actions and physiological changes are not a direct consequence of emotions, but part of the atomic units that construct them (Feldman Barrett, 2006).

Doubting also the universality of emotion expressions, this tradition advocates for a situational variability of emotions (Allport, 1924; Landis, 1924; Klineberg, 1940) which are learned from within a culture, corroborated for instance by the difference in the emotion vocabulary of different languages (Wierzbicka, 1999).

[1] Also referred to as “polarity” (Mohammad, 2016).

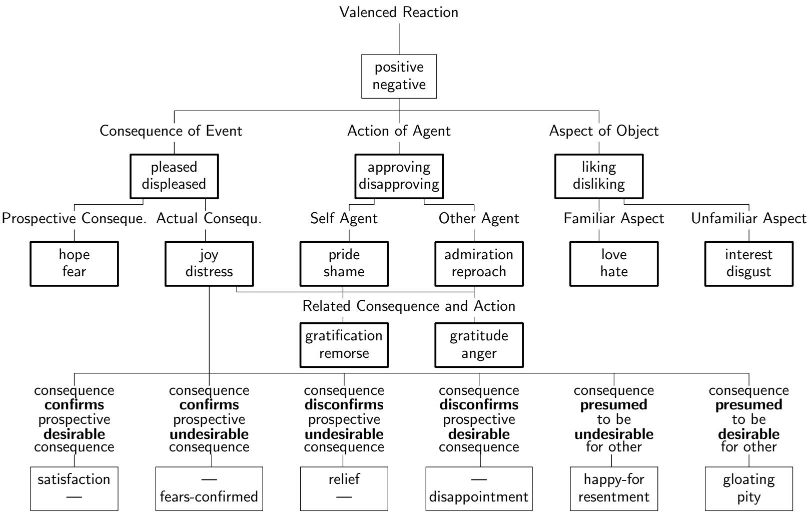


Figure 2.2: The OCC model, drawn after the depiction of Steunebrink et al. (2009, Figure 2).

1.1.3 Evaluative

The motivational and feeling traditions differ in their proposals of the elements that partake in an affective experience, their universality and direction of causality. Still, they concur that there exist some diagnostics of emotions (Scarantino, 2016). A starting cause is usually there (e.g., a fact happens); it is evaluated by its experiencers; and it sparks in them some concrete effects, like changes in their voice and posture. The evaluative tradition presents these observations as well, packing them into a perspective that relates emotions with sets of evaluations (Moors et al., 2013; Scherer, 1984).

To this tradition belongs the OCC model (Ortony et al., 1988), named after its authors Ortony, Clore and Collins. The model is reported in Figure 2.2. It formalizes the cognitive coordinates that rule more than 20 emotion phenomena (shown in the bold boxes of the figure) within a diagram.[2] The structure stems from three emotion-eliciting conditions,

[2] It is not uncommon to find discussions on the OCC model in terms of a hierarchy. Steunebrink et al. (2009) refer to it as an inheritance hierarchy, where each emotion "is like its parent type plus some specialization". For instance, displeased (in the left branch) is a negatively valenced reaction to an event consequence, distress adds on top of that a focus on the actual consequences for oneself.

namely, consequences of events, agents' actions and aspects of objects. The three branches further split to represent how people appraise events, agents and objects. Events are assessed with respect to one's goals, agents are evaluated against norms and standards, and objects bring into play representations of tastes and attitudes. Importantly, these assessments happen along some binary criteria, for instance desirability–undesirability, and they combine like logical functions. The OCC spells out ways in which specific evaluations are conducted one after the other (e.g., if a condition holds, a certain reaction follows). A given instantiation of evaluations, which can be traced in the diagram, fires an emotion deterministically. For example, the liking of an object sparks love.

Similar to other approaches in the tradition, the OCC model differentiates emotions with respect to their situational meanings. Yet, it sees them as simple descriptive structures of prototypical situations (Clore and Ortony, 2013), while other theorists qualify them in terms of a process, starting from an elaborate analysis of the cognitive changes that they involve.

At the core of the component process approach lies the idea that an emotion activates many resources useful to face salient circumstances. In the words of Scherer et al. (2001b), it is

> "an episode of interrelated, synchronized changes in the states of [...] the five organismic subsystems in response to the evaluation of a [...] stimulus-event as relevant to major concerns of the organism".

The five subsystems are cognitive, neurophysiological and motivational components (respectively, an evaluation, some bodily symptoms and action tendencies), as well as motor (facial and vocal) expressions, and subjective feelings (the perceived emotional experience, as something that feels either good or bad).

In practice, when a stimulus presents itself to an individual, it is assessed against the individual's goals, beliefs and desires. The changes in the state of the cognitive (i.e., information processing) component consist in weighting the importance of the situation. Does it hamper my goals? Can I predict what will happen next? Do I care about it? Two people with different goals, cultures and beliefs might produce different evaluations of a given circumstance, depending on personal features such as subjective values, motivational states and contextual pressures (Scherer et al., 2010).

These evaluations are immediate **appraisals** of situations (Arnold, 1960) which mobilize a *process* involving the four other components: bodily symptoms (e.g., heart beating faster), facial and vocal expressions (e.g. exulting loudly), action dispositions (e.g., jumping) and a feeling (e.g., feeling good). In this sense, event evaluations are complemented by multiple and specific effects. If I win the competition, I might smile and feel a pleasant sensation because winning supports my well-being; for the opposite reason, my opponent might not have the same reaction. Evaluations also determine the emotion that one experiences, so much so that they can be deemed as *causing* emotions (Scherer, 2005) or as *constituting* them tout court (e.g., in in Ellsworth and Smith (1988) appraisals are themselves emotions). For instance, fear emerges when an event is unforeseen, unpleasant and contrary to one's goal (Mortillaro et al., 2012).

The criteria that humans use to assess a situation are in principle countless, but researchers in psychology have come up with a finite number thereof that switches on for emotion-eliciting events. According to Ellsworth and Smith, there are six emotion-related appraisals: *pleasantness* (how pleasant an event is, likely to be strong for joy, but not for disgust), *effort* (that an event can be expected to cause, high for anger and fear), *certainty* (of the experiencer is about what is happening, low in the context of hope or surprise), *attention* (the degree of focus that is devoted to the event, low, e.g., with boredom or disgust), *own responsibility* (how much responsibility the experiencer of the emotion holds for what has happened, high when feeling challenged or proud), and *own control* (how much control the experiencer feels to have over the situation, low in the case of anger). Ellsworth and Smith find these dimensions[3] to be powerful enough to differentiate 15 emotion categories (Smith and Ellsworth, 1985, Table 6).

Scherer and Fontaine (2013) propose a more high-level and structured approach. Figure 2.3 illustrates their multi-level sequential process, which comprises four appraisal objectives that unfold orderly over time. First, an event is gauged to establish the degree to which it affects the experiencer (Relevance) and its consequences affect the experiencers' goals (Implication). Then, it is assessed in terms of how well the experiencer can adjust to such consequences (Coping potential), and how the event stands in relation to her moral and ethical values (Normative Significance). Each objective is pursued with a series of checks. For instance, organisms scan the Relevance of the environment

[3] Henceforth, I use "appraisals" and "appraisal dimensions" interchangeably.

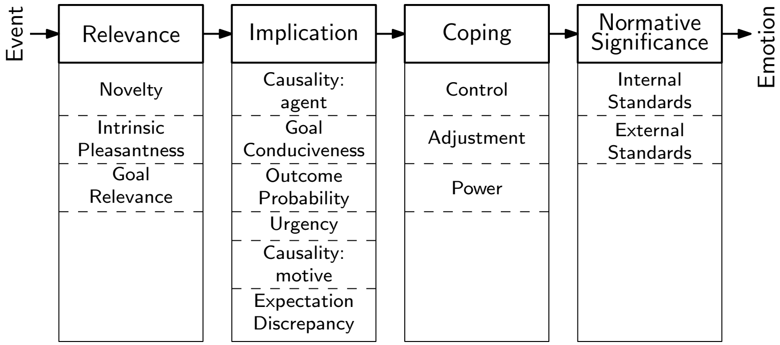


Figure 2.3: Sequence of appraisal criteria adapted from Sander et al. (2005) and Scherer and Fontaine (2013). High-level categories represent four appraisal objectives, with the item inside the dashed boxes corresponding to the relative checks.

by checking its novelty, which in turn determines whether the stimulus demands further examination; the Implication of the emotion stimulus instead is estimated by attributing the event to an agent, checking if it facilitates the achievement of goals, attempting to predict what outcomes are most likely to occur; the Coping potential of the self to adapt to such consequences is checked by appraising, e.g., who is in control of the situation; as for the Normative Significance, an event is evaluated against internal, personal values that deal with self-concepts and self-esteem, as well as shared values in the social and cultural environment to which the experiencer belongs. Each check can be broken down into one or many appraisal dimensions (not reported in the figure). As an example, the objective of Relevance encompasses the dimensions of *suddenness* (of the event), *familiarity* (of the experiencer with the event), and a degree of *event predictability* for the Novelty check, the event *pleasantness* and *unpleasantness* for the Intrinsic Pleasantness check, and *goal relevance* for the check concerning the salience of the event for one's goals.

1.2 An Argument for the Advantage of the Evaluative Tradition

We will see later that studies in computational emotion analysis face the choice to follow one of the three broad categories in which I re-grouped research from psychology, and to pick one specific theory within it (e.g., Ekman's or Plutchik's, Valence-Arousal-Dominance or Valence-Arousal). In this thesis, the component process theory from the evaluative tradition will turn out valuable for practical reasons, but there are also some theoretical advantages that motivate to use it in alternative to (or in combination with) other models.

The Advantage of Component Process Theories over the OCC Model. Like the OCC structure, also appraisal checks possess an underlying dimension of valence (Scherer et al., 2010). One always represents the result of a check as positive or negative for the organism: for intrinsic pleasantness, valence aligns with a concept of pleasure, for goal relevance with an idea of satisfaction, for coping potential, with a sense of power; it involves self or ethical worthiness in the case of internal and external standards compatibility, and the perceived predictability for novelty (with a positive valence being a balanced amount of novelty and unpredictability – otherwise a too sudden and unpredictable event could be dangerous, while a too familiar one could be boredom-inducing). Compared to the OCC structure, however, the component process theories account for more fine-grained evaluation criteria than good/bad distinctions.

Three Advantages of Appraisals over Discrete and Dimensional Models. Theories of emotions as motivations[4] and feelings[5] avail themselves of the importance of different emotional elements. Discrete perspectives offer a limited range of prototypes, which can be argued at odds with the actual affective states of people (Barrett, 1998), often complex and nuanced (e.g., being angrily disgusted). Dimensional ones not only face the problem of mapping continuous values onto emotion labels, but as pointed out by Smith and Ellsworth (1985), not all emotions can be distinguished solely by valence and arousal.

The evaluative tradition of appraisals pulls together aspects of the two. It retains the view of emotions emerging from some components or dimensions (cf. core affect) and spurring some physiological patterns (cf. Ekman), and in addition, it emphasizes the underlying cognitive mechanisms. This resolves the limitations of the others in at least three respects.

First, appraisals provide a structured differentiation among emo-

[4] From now on, I will refer to theories within the motivational tradition as the "discrete models" of emotion.

[5] Henceforth, I will refer to theories within the affect-based framework of the feeling tradition as the "dimensional models" of emotion.

tions. For instance, anger and fear are experienced when the cause of a negative event is attributed to external factors, whereas shame and guilt are felt, respectively, if the causes of an event are stable and uncontrollable traits of the self (e.g., being unable to focus) or controllable behaviors (e.g., not observing a speed limit) (Tracy and Robins, 2006). Consequently, theories putting attention on the cognitive nature of emotions have an advantage over discrete alternatives, because they overtly spell out the assessments that *justify* different emotion perspectives before the same stimulus (Mortillaro et al., 2012). They also tackle a biting problem for dimensional models, namely, the issue of differentiating emotions that have similar feelings (Bedford, 1957), such as indignation and annoyance.

Second, appraisals have an immediate relation to the way emotions are elicited. The origin of emotions is to be seen in the stimulus as appraised rather than the stimulus as such. In contrast to the Valence-Arousal-Dominance model, where it is left unclear if valence refers to a quality of the emotion stimulus or a quality of the feeling (Scherer, 2005), all appraisal dimensions are unambiguously event-directed. Similar to VAD, appraisals can be interpreted as a dimensional model of emotions, but as one that focuses on people's interaction with the surrounding environment.[6] Emotions are *about* particular things that are represented in particular ways (Scarantino, 2016); with their evaluation criteria, appraisal proponents account for this intentional feature of emotions in a transparent manner, which discrete and dimensional models fall short in formalizing.

Third, a critical issue of the feeling tradition is the thought that emotions have no causal effect on action: it is an instinct that drives action, which then causes bodily changes, eliciting in turn an emotion (James, 1890). Thinkers in the motivational tradition solve this problem. They explain that emotions do not "terminate" in an individual's body, but have the function of relating individuals to the stimulus that provoked them (e.g., via action readiness). By attending such a relationship, appraisal dimensions clarify the evaluations that motivate behaviors (e.g., flee, because the stimulus is dangerous).

[6] In an event-directed perspective, VAD concepts are implicit to some appraisals I will use in Chapter 3 (e.g., valence $\approx$ *pleasantness* $-$ *unpleasantness*, arousal $\approx$ *attention* $-$ *not consider*, dominance $\approx$ *own control*).

1.3 Emotion Recognition

Besides describing how emotions arise, research in psychology (and neuroscience) has explained how they are recognized. Understanding humans' ability to access emotions is an important theoretical asset. Since emotions have an interpersonal purpose, i.e., they modify the situation in which they occur (Parkinson et al., 2005), people interpret them in virtually all social transactions. That is relevant from a computational perspective because it is this ability that models should mimic.

Emotion recognition refers to an affective role-taking act. Similar to what Wispé (1986) calls empathy, it is an inference by which an observer discerns the internal state of an emoter, assuming her perspective but without necessarily feeling the same. The recognition applies to both verbal and non-verbal cues. Putting attention to the latter, Darwin (1872) was the first to sustain that emotions have universally recognizable facial expressions, a pathway for survival that communicates our needs to others. This idea led successive studies to code facial expressions. They formulated schemes of muscle movements as somewhat defined configurations from which observers infer that an emotion is ongoing (Ekman et al., 1980; Ekman and Friesen, 1978). Other research was more motivated by the fact that emotions are not always perfectly identified. It investigated the factors that influence affective-oriented face interpretations by considering elements on the side of the observer (e.g., age in Lawrence et al. (2015)) or features of emotion stimuli (e.g., color in Ikeda (2020)).

An example is Mancini et al. (2018). They pointed out that certain emotions are better inferred than others at a specific time in life (i.e., pre-adolescents identify happiness more easily than fear), and further, that there is a link between one's recognition performance and internal state. The relationship between the accuracy in seeing emotions and the emotion/mood of the person undertaking the task actually emerged in different works. For instance, Schmid and Schmid Mast (2010) revealed that people struggle more to grasp the emotion on another person's face if what they themselves feel is not aligned with that emotion. Vice versa, Niedenthal et al. (2000) discussed the effect of mood congruity between emoter and observer. Emotions are perceived as lasting longer if they match the state of the subject recognizing them. Negative mood, specifically, biases attention towards negative emotional cues in people with depressive disorders (Leppänen, 2006). This also happens in healthy subjects. When judging faces in the experiment of Bouhuys et al. (1995), participants in a negative mood perceived more sadness and less happiness compared to "happy" participants. In addition to that, Chepenik et al. (2007) noticed that negative moods are detrimental for people's emotion detection performance.

Also personality traits proved to influence the task: conscientiousness and openness are positively correlated with the ability to access emotions, in contrast to shyness and neuroticism (Hall et al., 2016). As for gender, Hoffmann et al. (2010) indicated that women and men perform comparably when dealing with highly expressive stimuli, but women seem to recognize less intense emotion expressions more accurately. The entanglement of intensity and emotion decoding appears in other modalities as well. In the experiment of Juslin and Laukka (2001), subjects exhibited higher emotion decoding accuracy with vocal stimuli that had strong emotion intensity. Lausen and Hammerschmidt (2020) found similar results, suggesting that certain prosodic cues that signal emotional bursts (e.g., pitch, loudness) have an effect on how well listeners infer emotions from speech.

In addition, Lausen and Hammerschmidt (2020) considered confidence, a metacognitive ability to reflect upon one's own (emotion recognition) answers. They took on a common procedure to investigate people's emotion abilities, by directly asking them to self-evaluate themselves (Schutte et al., 1998, i.a.). Humans can estimate the goodness of their inferences (i.e., How sure am I of the emotion I recognized?); thus, they can provide retrospective confidence ratings, which allow researchers to check how faithful a self-judgement is to the actual performance in a task (Bègue et al., 2018, i.a.). Lausen and Hammerschmidt (2020) found that not only prosodic cues but also people's correct emotion judgments related to higher confidence in their response. Kelly and Metcalfe (2011) drew a comparable conclusion with visual stimuli: individuals who did better in a facial emotion recognition task were more accurate in their metacognitive assessments.

Lastly, much research addressed the cultural implications of emotion recognition (Elfenbein and Ambady, 2002). While some scientists focused on the participants' disagreements due to their different cultural exposure (Elfenbein and Ambady, 2003a,b), Ekman (1972) was interested in their common understandings (i.e., to verify if people make similar inferences when seeing particular faces). He tapped on an external study showing that culture has no significant impact on emotion judgments (p. 242). The study considered the quality of the recognition performance within groups of Japanese and American subjects, who agreed on the emotions expressed in some pictures of faces with an accuracy of .79 and .86, respectively. However, by comparing the coders' decisions with the actual emotion felt by the depicted subjects, accuracy dropped to .57 and .62 (.5 being chance). That study has the merit to highlight that measures of agreement return different results depending on whether they are computed among judges of the emotions felt by others, or between the same judges and those "others" – an intuition that turns out relevant for this thesis.

2 Emotions in Language

Linguistic expressions of emotions are diverse. Borrowing the categorization of components from theories of appraisal, one can think of references to a stimulus event (e.g., "*I discovered my brother is a thief*"), to evaluations (e.g. "*that's unacceptable*"), to bodily cues (e.g., "*I was shaking!*"), to motor expressions (e.g., "*I ran away immediately*"), to motivational pulses (e.g., "*I wanted to forget all about it*"), and feeling descriptions (e.g., "*it was pleasant*"). These examples showcase two important aspects of emotions in language. First, none of the texts points out the writer's natural language interpretation of the affective state (e.g., anger), and yet all of them arguably suggest that she had one. Second, the same emotion episode can be signalled by verbalizing its different parts, like its effects or causes. Such examples represent the data of interest for computational emotion analysis, a field that has primarily preoccupied itself with the goal of automatic emotion recognition.

In NLP, machines that "recognize emotions" do not directly measure the affective states of people, which are only fully understandable by the direct subject through her embodied experience of physical changes, motor dispositions and so on. What the systems do is "inferring an emotional state from observations of emotional expressions and behavior, and through reasoning about an emotion-generating situation" (Picard, 2000). They exploit data as a function of emotions that any outsider could have access to.

While the data in question spans various modalities, like vision (Devillers et al., 2006; Williams et al., 2018) and audio (Munot and Nenkova, 2019; Bertero et al., 2016), this thesis only deals with text[7],

[7] Henceforth, by "emotion analysis" I mean computational emotion analysis in text.

which comes with challenges of its own. For example, emotion realizations in text are less coded than with faces; as opposed to videos, in written language there are no other cues than words. Texts also have appealing advantages. The vast amount of digitalized productions that currently exist makes it easy to obtain spontaneous expressions of written emotions, as opposed to, e.g. speech ones, which are often acted out (Sidorov et al., 2014, i.a.).

The problem of emotion analysis in text is formulated as a classification or a regression task. Textual units (documents, paragraphs, sentences, words) must be mapped to labels (i.e., the emotion names that a text can be associated with) or to numeric intervals (indicating the extent to which some emotion features can be inferred from text, like emotion intensity or VAD values). These outputs of computational models correspond to emotion estimates from a predefined text understanding perspective. They should approximate the emotion that writers felt when producing a text (Balabantaray et al., 2012), one applicable to an entity they name (Kim and Klinger, 2019), or one that the text elicits in the readers (Li et al., 2016b). Irrespective of their text understanding perspective, studies need a system of ground rules about the phenomenon they research in text. They usually look for one in psychology. It follows that a twofold partition separates the field: the type of used resources – either labeled corpora or lexicons, which provide emotion information at the level of different textual units, and the emotion theory selected as a reference.

Here, I summarize previous work on emotion recognition, following a common workflow of emotion analysis: as the resources on which the task is learned involve an annotation effort conducted by humans, I first discuss how data is labeled and the agreement is computed among annotators; then, I zoom in the resulting datasets and the automatic tasks learnt on them. Adapting the organization of Scarantino (2016, p. 8) to computational linguistics, I show how different studies match either the *feeling tradition*, the *motivational tradition*, or the *evaluative tradition*. While reviewing them, I make a distinction between works that leverage corpora and those that use lexicons.

2.1 Human Emotion Recognition

The creation of resources for machines to learn the tasks of emotion analysis requires two steps, a standard practice also for other fields. First is data annotation conducted by humans, where (mainly) already-existing texts are labeled by multiple coders independently, undertaking an emotion recognition task (Section 2.1.1). Then comes data curation, in which the judgments are analyzed to observe their agreement (Section 2.1.2).

2.1.1 Annotating Data: What and How

The related task of sentiment analysis, aimed at detecting the polarity of texts, can count on a range of corpora and dictionaries for multiple languages (Cieliebak et al., 2017; Galeshchuk et al., 2019; Chen and Skiena, 2014; Barnes et al., 2022). The situation is less fortunate for emotion analysis. Emotion annotation is slower and more subjective (Schuff et al., 2017). Plus, resources developed on a specific textual domain can turn out unusable for another (Bostan and Klinger, 2018), because a domain mismatch often implies a mismatch in the ways emotions are communicated.

The discrepancy between resources is also due to their compliance with alternative theoretical emotion models. As a matter of fact, the lack of a consensus about emotions in psychology has a clear implication for emotion analysis, which is left to pick an appropriate theory arbitrarily. The rationale that guides such a choice is usually pragmatic. It takes into consideration if the emotions documented by a theory are a good fit for the textual domain at hand and the emotion expressions it contains.[8] In turn, the selected theory affects how the data is annotated (e.g., Ekman's basic emotions, VAD scores) and what task can be performed on it automatically (e.g., classification, regression). To date, emotion analysis has mostly leaned on the motivational and the feeling streams of thought, specifically used for their discrete and dimensional models.

Motivational. Theories of basic emotions provide a foundational argument for affective computing: different emotions can be clearly grasped not only on faces but also when the communication channel is text. This is the main notion that computational studies borrow from discrete emotion models in psychology (e.g., those from Ekman, 1992; Plutchik, 2001), although the motivational tradition offers a much more varied picture (e.g., universality, emotions as "programs" that activate changes, retrievals of memories and expectations (Ekman and Cordaro,

[8] Not every label that NLP refers to as emotion is considered as such in psychology. For instance, Kim and Klinger (2019) characterize relationships between fictional characters by their affective undertone, but this, strictly speaking, violates the short-lived quality of emotion episodes (Scherer, 2005).

2011)). Hence, emotions occur through the contribution of multiple components, but this idea has received little attention among studies of discrete emotions in text. Their link to psychology lies only at the level of the task. Annotators associate discrete labels like *anger*, *disgust*, *sadness* to words or larger pieces of text.

Feeling. A wave of studies based on affect also exists in NLP, aligned with the constructivist conception that an emotion is like a "blend of hedonic (pleasure–displeasure) and arousal (sleepy–activated) values" (Russell, 2003), and that the changes along those continuous dimensions can then be mapped into discrete affective categories. This translates into the formulation of emotion annotation as a task that scores texts with respect to (all or a subset of) the continuous dimensions of Valence-Arousal-Dominance.

Evaluative. Automatic emotion recognition tasks have been explored in the wake of the evaluative tradition, but annotation projects that aim at populating textual corpora with evaluative dimensions are rare.[9] Such projects have been carried out in psychology, by means of self-reports of events associated with emotion labels and ratings for multiple appraisal dimensions.

With discrete models, coders are tasked to choose one or many emotion labels out of a predefined selection. Continuous labels of dimensional models (or intensity values for emotion categories) are rated on Likert scales or via best-worst scaling – a strategy in which items are arranged in tuples, they are compared to determine which ones are the most and the least representative of a certain variable (e.g., highest/lowest valence), and then are assigned a score for such variable by aggregating the comparisons (Louviere et al., 2015). These annotations are typically

[9] A stronger connection to this tradition can be found in general linguistics, where the discussion of Martin and White (2003) about the language of evaluations constitutes a full-fledged theory of appraisal in text. Tapping into the framework of Systemic Functional Linguistics, they analyze how interpersonal meaning is constructed in evaluation making. They provide a treatment of the linguistic possibilities with which writers share assessments and stances, and have considered at the same time how readers interpret them. While recognizing that the evaluative discourse permeates countless linguistic constructions, the analysis sharpens on a handful of mechanisms like appreciation, graduation (e.g., grading phenomena, amplifying feelings) and normative judgments. Emotions are a stratum of interpersonal meaning that is sourced through these mechanisms – more concretely, through mentioned qualities of entities, through modal adjuncts that reflect the position of writers towards an event (e.g., "*sadly, ...*"), through the communication of behavioral processes (e.g., "*he smiled at him*"), as well as mental (e.g., "*he liked him*") and relational ones (e.g., "*he felt happy at him*").

provided by external readers (≈ observers), while the texts' authors are rarely treated as annotators in an active manner. Indeed, resources containing information about their emotions tend to be constructed via self-labeling strategies, by leveraging hashtags (Mohammad, 2012), emojis (Felbo et al., 2017), and emotion-loaded phrases present in the text (Klinger et al., 2018) as proxies of emotions. Often, they are not re-annotated by external people (Lykousas et al., 2019).

Readers are required to assume a specific text understanding perspective, which corresponds to what the learning algorithms will be designed to do, and which depends on the domain of interest (e.g., in social media, the emotions/dimensions to be labeled could be those presumably intended by the writers; with poetry and news, those that the coders themselves are evoked; in literary texts, those associated with a character). Across these different scenarios, the annotation rubric can develop in one of two directions. Researchers recruit (many) laypeople or (a few) trained coders, and accordingly they conduct crowdsourcing campaigns or in-lab studies. Crowdsourcing allows compiling large data sources. Being labeled by naïve judges, they might turn out noisy and cause poor performance for the systems that learn the task from such data (Wauthier and Jordan, 2011). Instead, trained annotators are more reliable, but they can afford to cover smaller data, and this makes empirical observations difficult to draw.

There is in short a tradeoff between quality and quantity. However, this tradeoff is less pronounced than in fields focusing on predominantly linguistic phenomena. One can assume that emotion coders, although untrained, are familiar with the phenomenon they are called to annotate. Besides, emotions are subjective, and therefore expert judges can incur extreme disagreements all the same: section 1 discussed the factors entangled in the recognition of emotions beyond text (e.g., the emotion observers' age and current internal state), which correlate with the variability of judgments provided by people annotating the same stimulus; although unproven for all of the factors investigated in psychology for various stimulus modalities, it is reasonable to think that many of those apply to the linguistic domain as well.

2.1.2 Computing Agreement

As a next data creation step, the judgments are compared via measures of inter-annotator agreement (IAA). That is typically quantified among the readers. The agreement between readers and writers, whose judgments are rare to find, remains rather overlooked.

The text understanding perspective that the readers assume can be decisive for the annotation quality. Buechel and Hahn (2017b) provided evidence that readers who infer emotions from text by attempting to assume the writers' perspective achieve higher inter-annotator agreement than those who report their personal reactions. Other factors that can determine (dis)agreements are the annotators' characteristics. Mohammad (2018) analyzed their impact on VAD judgments, in line with the multiple works in psychology that delve into the observers' personal information to better understand their annotation performance. A significant relationship emerged between the task agreement and the demographic (age, gender) and personality traits (agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience) of the people who conducted it.

The calculation of IAA is influenced by the adopted emotion schema. With continuous annotations, agreement is assessed with measures of correlation or distance, like Pearson's r or Spearman's ρ, root mean square error (RMSE) and mean absolute error (Strapparava and Mihalcea, 2007; Yu et al., 2016). Solutions for the comparison of annotations of discrete categories are Cohen's κ (1960), for pairs of annotators, and its generalization to multiple coders, Fleiss' κ (1971). Since they quantify agreement for annotations where an item is assigned to one category, in multi-class classification problems it is common to calculate κ across all classes; in multi-label problems, that is done for each class separately. Previous work showed that the agreement between annotators in emotion analysis is limited in comparison to other NLP tasks, and that it varies substantially based on the textual domain of focus. Table 2.1 reports example Cohen's κ results achieved in different domains.

Cohen's κ is formally defined as $\frac{p_o - p_e}{1 - p_e}$. The term p_o is the observed probability of agreement; p_e is the agreement one can expect assuming that a random assignment of the texts into categories is guided by prior distributions unique to the coders, measurable from the distribution of labels that they assigned. Considering p_e is advantageous because it makes κ a chance-corrected coefficient. However, with skewed label distributions, chance agreement increases, penalizing the resulting score (Cicchetti and Feinstein, 1990).

Some authors overcome the issue by employing measures that are primarily employed as classification measures, like accuracy $= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$ and $\text{F}_1 = \frac{\text{TP}}{TP+\frac{1}{2}(\text{FP}+\text{FN})}$ (TP being the count of true positives, FP of false positives, TN of true negatives, and FN of false

Publication	Min κ	Max κ	Domain	Perspective
Aman and Szpakowicz (2007)	.60	.79	blogs	text
Alm et al. (2005)	.24	.51	tales	text
Haider et al. (2020)	.50	.84	poems	reader
Schuff et al. (2017)	.08	.57	tweets	unspecified
Kim and Klinger (2018)	.07	.35	literature	text
Štajner (2021)	.33	.55	tweets	writer
Volkova et al. (2010)	.34	.62	fairy tales	reader

Table 2.1: Examples of Cohen's κ across studies, textual domains, and text understanding perspectives. All use English corpora and discrete emotion models. "Min/Max κ" are variations across emotions. The labeled emotions are the affective profile of the text or an entity in it ("text"), the reader's reaction to the text ("reader"), or the writer's emotion inferred by the readers ("writer").

negatives). These can quantify agreement by treating the decision of one annotator as a gold standard and the other as a prediction (Štajner, 2021; Kim and Klinger, 2018). Accuracy and F_1 have a twofold advantage. Not only do they overcome the drawback of κ (i.e., the proneness to underestimating agreement). Since they are employed to evaluate classifiers, they also allow to directly compare human and automatic classification performances, and to take the former as an upper bound for the computational models (some will be detailed in the next section). For instance, if annotators agree with one another or with a pre-existing emotion label of a text only to a limited extent, models showing analogous performance are acceptable.

Agreement is informative at other levels as well. First, it is an indicator of the difficulty to identify some labels. Those characterized by the best agreement are the easiest to recognize, vice versa for the labels with the worst agreement (Bobicev and Sokolova, 2017). Second, IAA can help researchers determine if any filtering strategy should be applied on the judgments to resolve ties (Bhowmick et al., 2008) or to decide on a final, adjudicated ground truth (e.g., the majority vote for a text). Furthermore, agreement can signal the quality of datasets created automatically. For instance, Mohammad (2012) gathered a corpus of tweets labeled with emotion hashtags. Such a strategy is noisy, because a hashtag might not correspond to the emotion of the tweet. Still, an emotion classifier reached similar results on the "self-labeled" data and on manually annotated texts (40.1 F_1), suggesting that the quality of the labels was comparable in the two settings.

2.2 Automatic Emotion Recognition

Once resources are annotated and the annotation quality has been observed, the next step in the emotion analysis workflow is automatic emotion recognition. In what follows, I introduce existing computational methods to address this task (Section 2.2.1), and illustrate how they have been applied on the available resources (Section 2.2.2).

2.2.1 Models

To review emotion recognition methods, I divide them into two major categories[10], shown in Figure 2.4. The first group encompasses "expert systems" that reason via bodies of prior knowledge (cf. Jackson, 1986). The data-driven group can solve the task without strictly needing external information at classification time. I describe them below. My objective is to give a high-level view over a few ideas that motivated research from the first steps of computational emotion analysis up to date. To contextualize my own research, I focus on the recognition task that the thesis uses, i.e., text classification, but similar techniques can be applied when dealing with regression goals.

Prior Domain Knowledge. Methods based on domain knowledge detect textual emotions through explicit rules, as well as information stored inside linguistic resources, such as emotion dictionaries and ontologies. An early classification strategy in the field consisted in keyword spotting (Strapparava and Valitutti, 2004; Dodds et al., 2011; Strapparava and Mihalcea, 2008; Neviarouskaya et al., 2009). It required looking for a match between the terms of an input text and those in the considered resource.

Dictionaries provide access to the prior (i.e. out of context) emotion of individual words, where affective associations are formalized as labels or scores. A compositional function is therefore required to build lexical emotion meanings up to that of larger pieces of texts. Strapparava and Valitutti (2004) and Dodds et al. (2011) did that by observing the frequency of the emotion-laden terms in the text. Neviarouskaya et al. (2009) used predefined rules that specify how to weight words

[10] Slight variations of this arrangement can be found in the literature. See for instance the survey of Hakak et al. (2017) and Kao et al. (2009).

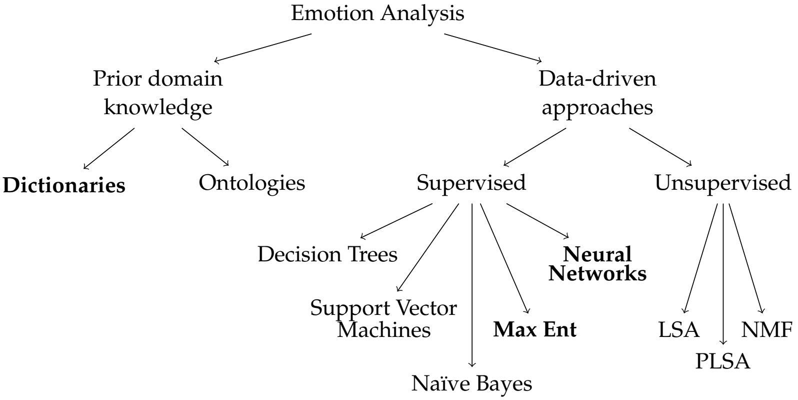


Figure 2.4: High-level overview of approaches to the problem of emotion recognition. Methods that will be presented in this thesis are in bold.

(or other linguistic constituents) that correspond to certain parts of speech. Shaikh et al. (2009) followed *if–then* rules motivated by the OCC model, which establish possible ways to synthesize different affective properties extracted from a text with the help of dictionaries. By way of example: if (Linguistic_Token_found_for_Joy(text) and No_Negation_Found(text)) or (valenced reaction = 'true'[11] & self reaction = 'pleased' and self presumption = 'desirable'[12]), 'joy' is true – i.e., the actor in the text feels joy if she is pleased about a desirable event.

Ontologies are other valuable sources of emotion information (Shivhare et al., 2015). An ontology is an explicit specification of a conceptualisation, i.e., a database that represents knowledge about a domain (in this case, emotions) in a declarative formalism (Gruber, 1995). It contains commonsense knowledge about entities, objects, classes thereof, as well as their properties and relations (Alfrjani et al., 2016), which can be modeled in a text for the purpose of classification. For instance, one can identify the action described in an input text (e.g., passing an exam), compute the similarity with the actions contained in the used emotion ontology, select the one with the highest semantic similarity to the text's, and use its emotion as the label for it (Balahur and Hermida, 2012).

[11] The valence of the sentence is positive.

[12] Both "self reaction" and "self presumption" are derived from the valence of an event.

These resource-lookup strategies have been successful in the field thanks to their transparency (Ling et al., 2006; Krcadinac et al., 2013, i.a.). They give researchers control over how the emotions of individual words combine to form the text's. On the downside, they compel to enunciate rules of composition for different linguistic constructions, in particular for negations and other linguistic modifiers that shift the meaning of words (Liu et al., 2003). Further, methods that align the content of texts with dictionaries typically do not account for the impact of word order: without appropriate treatment, they can lead to (incorrectly) assign the same emotion label to sentences with similar surface forms (e.g., "*I laughed at him*" vs. "*He laughed at me*") (Samonte et al., 2017). Ontologies are not unproblematic either. They fall short in considering the semantic relationship between a text's components that correspond to concepts in the ontology (e.g., actions) and the linguistic context in which they occur. Lastly, these resources might not necessarily fit all textual domains, because the same word can assume different connotations depending on the topic under discussion; in fact, in case their vocabulary has no overlap with the text in question, the emotion of the text cannot be captured at all.

In sum, prior knowledge-based solutions have a hard time in terms of generalization.

Data-driven Approaches. Data-driven classification methods are less prone to raising those issues. Rather than requiring hand-crafted rules and composition functions, they generalize to unseen texts a function previously *learned* on large corpora. In the field, the learning procedure has been mostly conducted in an unsupervised or supervised scenarios (Canales and Martínez-Barco, 2014) – respectively, one that does not involve a system to learn from labeled examples, and one that does.

Among the unsupervised approaches are techniques for dimensionality reduction and topic modeling such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999). Initially devised for information retrieval, they estimate the underlying latent semantic structure in the data. In practical terms, they can be used to map a matrix representing the terms in a document to a latent semantic space. Such a representation can then be leveraged to compute the semantic similarity between the texts and the vector corresponding to emotions keywords within the same space (Gill et al., 2008; Wang and Zheng, 2013; Calvo and Mac Kim, 2013).

The emotion classification literature is dominated by supervised classification approaches in machine learning, which train models on corpora associated with emotions at the level of one or many sentences. The making of these classifiers involves the optimization of multiple parameters with objective functions that encourage the labels predicted for the (training) examples to be as close as possible to the actual labels (Mohri et al., 2018). The models learned this way can then be used to classify new texts.

Word n-grams, character n-grams, punctuation, negations are treated as lexical features that link what one observes in the data with the emotion classes one wants to predict: given an input text whose emotion is unknown, classifiers are fed with representations of textual features that transform sequences of words into numerical vectors, and they output a probability distribution over a predefined set of emotion labels, returning the one (or a certain number of those) with the highest probability score(s).

The representation of lexical features can consist of word counts, like bag-of-words, that can be fed to models like decision trees (Alzu'bi et al., 2019) and SVM (e.g., in Balabantaray et al., 2012; Roberts et al., 2012), Naïve Bayes classifiers and Maximum Entropy (Max Ent) classifiers. A Naïve Bayes computes posterior probabilities based on the distribution of words in a document (i.e., it employs the joint likelihood between the data and the classes), while Max Ent is trained to maximize the conditional likelihood of the data. Unlike Naïve Bayes, Max Ent makes no assumption about the independence of features used as input (Berger et al., 1996), and therefore has the potential to perform better when the conditional independence assumption (between words, n-grams, etc.) does not hold. I will exploit it in Chapter 3.

Word count representations are, however, inefficient. First, they are vectors whose length corresponds to the size of the vocabulary (Elman, 1990), and thus become computationally expensive as the vocabulary increases. Further, they give no way to capture semantic similarities between words. To overcome these issues, most works today rely on neural networks fed with high-dimensional, distributed meaning vectors. Often referred to as "word embeddings", these vectors allow to fit extremely large (training) sets of possible word meanings with a relatively compact number of parameters. They can capture both syntactic and semantic word similarities. This promotes the generalization capabilities of the models (Collobert et al., 2011).

Embeddings are usually learned by neural network language models on data sources orders of magnitude bigger than those available for emotion classification (e.g., containing billions of tokens (Pennington et al., 2014)), and by learning to accomplish a sentence completion task. The training procedure involves multiple layers trained via backpropagation of a gradient-based optimization algorithm that maximizes the log-likelihood of the training data (Sun et al., 2022). The training objective is to predict next words or tokens conditioned on a variable context length (n words before the target one).

Since corpora are labeled at larger linguistic chunks than words, word representations need to be aggregated. Multiple architectural principles have been proposed for this step, promptly increasing the complexity of the linguistic contexts that the models capture. In Recurrent Neural Networks (e.g., Mikolov et al., 2010), words are fed consecutively, one at a time. Using a memory vector as input and an internal state, they produce a prediction for the next time step. Due to connections that allow for a time delay (Pascanu et al., 2013), efficient gradient-based learning with Recurrent Neural Network is hampered by a vanishing gradient problem (Bengio et al., 1994): with long text sequences, the loss information that is backpropagated during training shrinks. Long Short-Term Memory (LSTM) architectures (e.g., Merity et al. (2018) and Howard and Ruder (2018), based on Hochreiter and Schmidhuber (1997)) have been implemented, which maintain an internal state that controls what information to retain after every input, and thus captures long-distance dependencies. Bi-Directional Long Short-Term Memory (BiLSTM) (Schuster and Paliwal, 1997) later improved LSTM-based models by considering past and future information at each point of the processing of an input, with the result that the sentential context is integrated into the word representations. I will use a BiLSTM in Chapter 5.

State-of-the-art language models moved away from step-by-step reading, focusing on entire sentences. Attention-based transformer architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021), are fully-connected networks that allow for non-sequential processing, and that encode information about word positions to capture dependencies. Their multiple attention heads further permit them to focus on specific parts of an input, producing context-sensitive representations. Differently from what would happen in a BiLSTM, small surface variations can create large variations in the sentence representations. For this reason, transformer-based models are promising to capture emotion differences between texts that largely overlap at the lexical level (taking back the example above, *"I laughed* *at him"* vs. *"He laughed at me"*). I will use and RoBERTa in Chapter 3, and BERT in Chapter 4 and Chapter 6.

The performance of language models is evaluated via perplexity, which quantifies, intuitively, how much surprise there is in reading the sequence of predicted words. The lower perplexity, the better the model (Fossum and Levy, 2012; Goodkind and Bicknell, 2018). Having generative (sentence completion) capabilities, language models are trained for generation purposes, by optimizing objective functions that concern specific subtasks – e.g., for neural machine translation, fluency (the naturalness of the predicted text, well captured by perplexity) and accuracy (how faithful to the input it is) (Bahdanau et al., 2015; Sutskever et al., 2014). For the task of classification, it is unusual to train models from scratch: the parameters of pretrained models can be fine-tuned on new training data labeled with the variable of interest, for the representations to become adequate for a specific classification goal (Ruder et al., 2019). That involves adding a softmax layer, an element-wise logistic function that is normalized such that all elements sum up to 1 (Bishop and Nasrabadi, 2006). In emotion classification, these scores represent a probability distribution over emotion classes.

More and more technologies are being developed in NLP that fit the goal of emotion-based text classification. Therefore, I will tackle the task with different architectures. I will fine-tune neural networks on corpora of various sizes, and I will opt for them when dealing with different textual domains, leveraging their stronger cross-domain generalization capabilities than probabilistic classifiers, such as Max Ent. The choice of various architectures in different chapters has had three reasons. First, the presence of similar models used for the same task in the field, that I could compare to. Second, the lack of convincing evidence that other models would have performed substantially better. Had I run the experiments with different architectures, the classification performance numbers could have turned out different, but they would hardly have altered the high-level findings. My research aims indeed at using classifiers as tools to foster linguistic understanding of emotions, not at comparing models. Lastly, as different models were employed at different points in time, they reflect my personal growth within the field.

2.2.2 The Psychological Traditions in Automatic Emotion Recognition

We can now consider how the models I reviewed have been put into practice in tandem with specific resources and with emotion theories from psychology. Depending on the data under consideration and the psychology-driven emotion schema that annotates it (one among the three discussed on Page 31), the task of automatic emotion recognition is instantiated in a specific manner.

Mirroring the annotation tasks that adopt discrete models, the goal of automatic emotion recognition studies that match the motivational tradition is to assign categories to words (Mohammad and Turney, 2013) or sentences (Felbo et al., 2017; Li et al., 2017; Schuff et al., 2017). For that, both dictionaries and corpora are available.

Works adopting a dimensional approach, which corresponds to the feeling tradition, face emotion recognition as the problem of predicting VA(D) scores (e.g., Yu et al., 2016; Buechel and Hahn, 2017a). Dimensional models present the advantage of formalizing differences between emotions with a handful of continuous emotion dimensions. Hence, the systems bypass the decision of picking one out of various states that are similar in respect to some dimensions (e.g., both sadness and anger are characterized by low valence), and that could equally hold for a given text. They only need to learn relations between valence, arousal and dominance. In consequence, they can account also for feelings that do not fall in the crisp separation between basic states. This task can then be followed by an optional VA(D)-to-discrete emotions mapping.

Lastly, a slice of emotion analysis has acquired from psychology more than labels or affect components to investigate in text. It has turned to the evaluative tradition to formalize the evaluations that stand beyond emotion episodes, as components that should help solve the same classification problem addressed with discrete emotion models. This research direction is far less explored than the other two. Its work on emotion recognition has predominantly been conducted with the tools of the OCC model, which uses logics-based representations. Appraisal dimensions have been mostly dismissed, despite the theoretical merits of the component process model listed in Section 1. The study of Balahur et al. (2012) is an exception. It achieved promising emotion classification results by leveraging appraisal-oriented notions: with the help of an ontology, they inferred characteristics (≈ appraised properties) of actors, actions, and objects from text.[13]

The remainder of this section expands on works on emotion analysis, divided by the psychological traditions and resources they focus on. While diverging in many respects, the motivational, feeling and evaluative streams of thought on emotions provide insights that share some commonalities. For instance, dimensional models are based on affect, but they allow for the mapping from continuous concepts to discrete ones, in line with basic models. This suggests that also the corresponding approaches in computational emotion analysis should not be thought of in conflict, but rather as complementary.

Motivational. Evidence has shown that every language has some lexical items to pinpoint emotional experiences (Wierzbicka, 1995), and that a large part of its vocabulary can be described in terms of emotion meanings (Clore et al., 1987; Hobbs and Gordon, 2011). Linking language and emotions via lexicons thus represented a straightforward passage from theory to practice for affective computing. Strapparava and Valitutti (2004) did that with a semi-automatic procedure that resulted in WordNet Affect. They equipped the existing thesaurus WordNet (Fellbaum, 1998) with a series of affect-related concepts, like *emotion, cognitive state,* and *mood*, that characterize words, in turn sorted into annotated emotional synsets (groups of words expressing similar concepts). Items from WordNet Affect were later included in other resources. One is Affect (Neviarouskaya et al., 2007), a dictionary of terms that fit instant messaging texts, which contains, e.g., emotion-related emoticons (e.g., “ =^_^= ” stands for a blushing reaction) and abbreviations (e.g., “BL”, belly laughing) assigned emotion categories and intensities. Another is the NRC Emotion Lexicon, which was crowdsourced and is a higher-coverage dictionary, including also common English bigrams (Mohammad and Turney, 2013).

Lexicons are handy for classification problems. As a naïve baseline, Strapparava and Mihalcea (2008) checked for the presence of the words from WordNet Affect in newspaper headlines, to determine the emotion of the latter. The proposal of Neviarouskaya et al. (2009) took a step further to account for the compositionality of sentence-level emo-

[13] Neither the ontology (EmotiNet) nor the emotion classification step focused on the rich set of event appraisals established in the literature. More than being a resource of appraisal information, EmotiNet is inspired by the idea that emotions are *reactions* to real-life contexts. It allows to model the content of texts as action chains, “changes produced by or occurring to an agent related to the state of a physical or emotional object” (Balahur et al., 2012, p. 89), that take place within a context.

tions given the emotions of their sub-constituents. It used the resource Affect in a bottom-up, rule-based algorithm: first, it represented the emotions of words in terms of intensity; next, it applied pre-defined mathematical operations to quantify interactions between the intensities of different phrase constituents, for example to quantify how a modifier increases or decreases the emotion of an adjective; lastly, it decided on the emotion of the entire sentence, based on relations between subjects, verbs and objects (e.g., in some cases, the vector of maximum intensities from the subject, verb, and object, corresponds to the emotions of the whole text).

On the other hand are models for discrete emotion predictions on corpora. They are by and large standard text classification approaches (Sailunaz et al., 2018). Early supervised approaches opted for linear classifiers. Strapparava and Mihalcea (2008) and Alm et al. (2005) respectively worked with Naïve Bayes and the learning architecture SNoW, fed with linguistic features (adjectives, verbs, exclamation marks, parts of speech, punctuation marks). Later, neural models proved to achieve superior performance: comparing the classification results obtained via Maximum Entropy, SVM, CNN, Long Short-Term Memory, Bi-Directional Long Short-Term Memory, Schuff et al. (2017) found evidence that neural networks perform the best across the board of Plutchik's eight fundamental emotions.

Recent shared tasks (Klinger et al., 2018; Mohammad et al., 2018; Mohammad and Bravo-Marquez, 2017) indicate a shift of attention toward transfer learning, which exploits the availability of large resources related in content to emotions, to then address the recognition task on smaller data, less suitable for training. Among those are Felbo et al. (2017), who used emoji representations for pretraining a model for emotion classification, and Cevher et al. (2019), who pretrained a neural network on existing emotion corpora followed by fine-tuning on a narrow domain with considerably less training data.

Nowadays, ready-to-use corpora span many domains, like stories (Alm et al., 2005), news headlines (Strapparava and Mihalcea, 2007), songs lyrics (Mihalcea and Strapparava, 2012), conversations (Li et al., 2017; Poria et al., 2019), and literary texts (Kim et al., 2017). An established resource is the one of Aman and Szpakowicz (2007), with sentence-level annotations for more than 5k blog texts. Other widely adopted corpora were compiled from social media as well. TEC (Mohammad, 2012) contains ≈21k tweets extracted by leveraging hashtags that correspond to the six emotion classes of Ekman (1992) (e.g., #anger), while IEST (Klinger et al., 2018) reaches almost 200k posts obtained by polling the Twitter API for synonyms of the Ekman's categories. The solution of building resources by harnessing emotion-related hashtags has been explored also in a considerable number of other data creation efforts (Roberts et al., 2012; Wang et al., 2012, i.a.).

The majority of these resources limit their labels to a set of four to eight fundamental emotions. Only a handful uses more. The corpora released by Abdul-Mageed and Ungar (2017) and Demszky et al. (2020) respectively contain tweets with all 24 emotions present in Plutchik's wheel, and Reddit comments associated with 27 emotion categories. EMPATHETICDIALOGUES supplies crowdsourced descriptions of emotional situations spanning 32 emotion labels.

Feeling. Similar to the discrete approach, dimensional lexicons contain words, but associated with VAD values. Hence, the scores of words present in a sentence can be used to represent that sentence in the affective space, by averaging the individual VAD scores (Calvo and Mac Kim, 2013) or, to cast lexical-based emotion recognition as a regression problem, by leveraging bag-of-words features (Buechel and Hahn, 2016).

For a long time, ANEW (Bradley and Lang, 1999) constituted the benchmark for lexicon-based research on emotion recognition. ANEW is a collection of words in English, with corresponding emotion norms rated by multiple readers. Each word w is associated to three scores averaged among annotators, that constitute the coordinates for w's representation in the space of affect (i.e., $w = (valence, arousal, dominance)$). Initially containing around 1k entries, ANEW has later been extended with different strategies. Warriner et al. (2013) used crowdsourcing, showing that non-expert judges produce annotations highly correlated with the original ratings of ANEW for all VAD dimensions. Bestgen and Vincze (2012) assigned emotion norms to new words ($\approx$17k) without the help of human-based experiments, but with a bootstrapping procedure. Their algorithm combined the use of a corpus, of ANEW, and of Latent Semantic Analysis: it estimated VAD norms based on the proximity in a semantic space of an unrated word with others, by averaging the (known) norms of the latter.

Later lexicon creation efforts proceeded similarly. They automatically assigned VAD scores depending on a term's semantic similarity with others endowed with manual annotations (Köper et al., 2017; Buechel et al., 2016). To date, word-level resources are available for both English (Bradley and Lang, 1999; Warriner et al., 2013; Mohammad, 2018) and other languages (e.g., Buechel et al. (2020) created lexicons for 91 languages, including Korean, Slovak, Icelandic, Hindi).

Only a few corpora are annotated at the sentence level with (at least a subset of) VAD information. Bradley and Lang (2007) created ANET, the corpus counterpart of ANEW, with affective norms for brief English texts. Other resources cover English as well (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017a,b), and some exist for Mandarin (Yu et al., 2016), Polish (Imbir, 2017), Spanish and Arabic (Mohammad et al., 2018).

With corpora, research investigated how the valence and arousal that emerge from text co-vary with certain attributes of the writers, such as age and gender (Preoţiuc-Pietro et al., 2016). It also attempted to close the gap between dimensional and discrete emotion models. Some authors dealt with it as a post-processing of a text's VAD values. For example, Calvo and Mac Kim (2013) measured the similarity between the vector of the text represented in the VAD space and the emotional categories represented in it as centroids of several keywords. They turned the VAD scores of the text into the discrete class with the highest similarity. Park et al. (2021) integrated the two emotion schemes earlier in the modeling stages, with a joint model that predicts fine-grained emotion categories together with continuous values of VAD. They did so using pretrained RoBERTa (Zhuang et al., 2021), fine-tuned with earth movers distance (Rubner et al., 2000) as a loss function to perform classification. Related approaches learned multiple emotion models at once, showing that a multi-task learning of discrete categories and VAD scores can benefit both subtasks (Akhtar et al., 2019; Mukherjee et al., 2021). In this vein, Buechel et al. (2021) defined a unified model for a shared latent representation of emotions, which is independent of the language of the text, the used emotion model, and corresponding emotion labels.

Evaluative. The dictionaries of the (computational) evaluative stream of research were originally meant for sentiment analysis. Sentiment lexicons are useful to adapt the building blocks of the OCC in the textual domain: like the OCC model, they involve polar (positive/negative) judgments. Both Shaikh et al. (2009) and Udochukwu and He (2015) used lexicons of valence scores to measure the variables from the theory of Clore and Ortony (2013): pleasantness, desirability, compatibility with goals and standards were represented with lexicons that associate objects and events with positive or negative scores, a confirmation status was associated with the tense of the text, and causality was modeled via semantic and dependency parsing. These variables were then combined with logical rules to infer an emotion category for the text. Although transparent, such an approach treats evaluations in isolation, focusing solely on those that have a textual realization; consequently, the emotion classification task is reduced to a deterministic decision that disregards probability distributions across multiple appraisal variables.

As for corpora, a textual dataset based on appraisal theories was compiled with a large in-lab data collection campaign in psychology. ISEAR (Scherer and Wallbott, 1994) contains self-reports collected in 37 different countries. It was built by asking university students to recall an emotion-inducing event, and to fill in a questionnaire where they described it, specified its characteristics, the feelings it evoked, and the associated physical symptoms. As a result, ISEAR provides 7665 instances directly labeled by the emotion experiencers themselves with one seven labels (i.e., *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*), and with a large number of ratings for many appraisal dimensions. The resource has been popular in emotion analysis (Danisman and Alpkocak, 2008; Pool and Nissim, 2016; Das and Bandyopadhyay, 2011; Boldrini et al., 2010) but not to exploit its appraisal-related annotations.

In summary, the discussion so far established multiple notions that will be handed on to later chapters. I summarized the constructs of different emotion models in psychology, insisting on **the link between emotions and events**, which is the core of appraisal theories. I further highlighted two insights that will turn out crucial for some design decisions in the dissertation. First, multiple **subjective factors play a role in emotion recognition from faces** and other stimuli. Second, **comparing the task performance between different groups of people** (e.g., among observers or between observers and experiencers) **can drastically affect a researcher's conclusions about inter-annotator agreement**. The discussion of emotions in language followed similar lines. By reviewing the basics of data annotation (i.e., the sequence of label collection and computation of agreement), I showed how different psychological models of emotions correspond to specific annotation schemata and allow instantiating the task of emotion recognition in different ways. Lastly, I introduced the available resources and methods to accomplish the task.

Regarding the loose theoretical outset of computational works, I pointed out that **appraisal theories, although theoretically convenient,**

have not been fully explored for emotion recognition. This defines a wide gap in the field.

3 Emotions and Style

Besides text classification, emotions have supported the task of style transfer, for which their study is combined with that of verbal styles. Style transfer is researched from various disciplines actually, fueled by the idea that communication, in many of its kinds, has two sides to it. Be it in language, visual arts, music, or other expressive channels, the things that people produce have a *content* (what is to be conveyed) and a *style* (how that is done). To use a linguistic example, these concepts are evident in the Shakespearean verses "*By the pricking of my thumbs, Something wicked this way comes*" (Macbeth, Act 4, Scene 1). Content (i.e., the foreseeing of an evil future) is a semantic nucleus, and it is encoded in a slant rhyme with peculiar rhythm and unusual vocabulary choices, that is, style, which corresponds to the form shaping a core piece of information, and which collocates it under some distinctive communicative categories (e.g., a poem, an old variety of English).

From a computational perspective, the content–style dichotomy turns out extremely interesting. By treating the two terms as independent variables, one can automatically produce content that is styled in a controlled manner. Several studies have indeed aimed at creating content from scratch, like texts (Gatt and Krahmer, 2018), images (Wu et al., 2017) and music (Briot et al., 2020), that feature the desired style. Style transfer has approached the generation problem to re-style pieces of content that already exist.

This practice is pervasive among humans in non-computational contexts. It can be observed any time they give an inventive twist to their verbal and artistic expressions (e.g., when conveying a literal gist through a metaphor or when painting by imitating Van Gogh's singular brush strokes). Correspondingly, the style transfer rationale is: if style and content are two and separate, one can be automatically modified and the other kept unaltered. The attempt has been successful in the field of computer vision, where research has been modifying the styles of images (Gatys et al., 2016). Following its footsteps, NLP has set a similar goal but with language, defining the task of textual style transfer as the generation of style-controlled textual paraphrases.

The possibility to automatically produce stylistic variations in language can be imagined handy in many practical scenarios. To name just a few: a transfer from technical to phatic jargon can adjust the level of literacy of texts, making them accessible to a broader audience; the task has the potential to assist automatic writing (e.g., to support non-native speakers produce polite responses, as they might ignore some semantic nuances of the target language). Moreover, style transfer is appealing from a theoretical perspective. Its manipulation of stylistic markers offers different conditions of investigation to explain how readers decide about the membership of a text into a certain linguistic category. Ultimately, it can help elucidate what style is.

In this section, I focus on textual style transfer carried out with the style "emotion". After defining the general transfer problem and summarizing the typically-applied methods and evaluation metrics (Section 3.1), I organize the styles present in the literature within a hierarchy (Section 3.2). The hierarchy develops following some communicative properties of the styles tackled so far in the field, and by that, it emphasizes their (dis)similarities. I show how emotions fit into it. Next, I provide an overview of past work, and I highlight the challenges that emotions pose when deemed as a style to transfer (Section 3.3).

3.1 Style-Aware Paraphrasing

Textual style transfer aims at modifying the style of texts while maintaining their initial content (i.e., their meaning). It represents an effort of conditioned language generation but differs from this broader task fundamentally. The latter creates text and imposes constraints over its stylistic characteristics alone. Style transfer starts from a given input text, and it introduces constraints on both style, which has to be different between input and output, and content, which has to be similar between the two – for some definition of "similar". The task can be considered a variant of automatic paraphrase generation (Zhou and Bhat, 2021). Both have the objective to re-phrase and yet preserve the input meaning, but only style transfer conditions paraphrases toward a target style.

Current publications use different evaluation metrics and generation methods to reach that goal. However, as I detail below, works present some common points.

Task Definition. Deep learning-based style transfer endeavors require the learning of $p(t' \mid s, t)$. A text t' has to be produced given the input t and a desired stylistic attribute s, where s indicates either the presence or the absence of such an attribute[14] with respect to t. For example, if t is written in a formal language, like the sentence "*Please, let us know of your needs*", then s may represent the opposite (i.e., *in*formality), requiring t' to shift toward a more casual tone, such as "*What do you want?*". Hence, a successful style transfer output checks three desiderata. It **exhibits a different stylistic attribute** than the source text t, it **preserves its content**, and it **reads as a human production** (Mir et al., 2019).

Evaluation. Evaluation metrics aim at computing how well the produced paraphrases meet the criteria of content preservation, transfer accuracy/intensity and readability.

Content preservation, i.e., the extent to which an output retains the content of the input, is usually gauged with measures of accuracy for machine translation. These compute the overlap between the words of the generation system and some reference texts, under the assumption that the two should share much lexical material. Among them are BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), often complemented with ROUGE (Lin, 2004), initially a measure for the quality of automatic summaries. Transfer accuracy, i.e., the efficacy of the models in varying stylistic attributes, is usually scored by classifiers: trained on a dataset characterized by the style in question, a classifier can tell if an output text has the target attribute or not. Last comes the readability or naturalness of the paraphrases. This is typically estimated with metrics for text fluency, such as the perplexity of language models, which indicate the degree to which the sequence of words in a paraphrase is predictable, hence grammatical.

A detailed discussion of the style transfer evaluation strategies is elaborated in Mir et al. (2019), Pang (2019), Briakou et al. (2021a) and Briakou et al. (2021b).

Methods. A style transfer pipeline usually comprises an encoder-decoder architecture inducing the target attribute on a latent representation of the input, either directly (Dai et al., 2019) or after the initial attribute has been stripped away (Cheng et al., 2020). Different frameworks have been formulated on top of this architecture, ranging from

[14] An "attribute" is the value (e.g., presence, absence, degree) that a specific style (e.g., formality) can take.

lexical substitutions (Wu et al., 2019b; Li et al., 2018) to adversarial techniques (Pang and Gimpel, 2019; Lai et al., 2019).

Surveys on the topic point out that the methodological choices of researchers are heavily limited by the data available for specific styles (Hu et al., 2020; Jin et al., 2022; Toshevska and Gievska, 2021). In the ideal scenario, a transfer system observes the linguistic realizations of the attributes of interest directly from a parallel corpus, which contains texts with a stylistic attribute on one side (e.g., formal texts) and paraphrases with a different attribute on the other (e.g., informal texts). When similar datasets are accessible, the task becomes a translation problem that maps one attribute into the other instead of operating a transformation from language to language (Xu et al., 2012; Rao and Tetreault, 2018). However, that is rarely the case because parallel resources for style transfer only exist in short supply (Jin et al., 2022). Resource paucity has thus triggered some attempts to synthesize parallel examples (e.g., Zhang et al., 2020b; Jin et al., 2019), and it has especially pushed the development of transfer strategies on mono-style corpora (e.g., John et al., 2019; Wu et al., 2019b; Fu et al., 2018; Cheng et al., 2020; Lin et al., 2020).

Mono-style resources (i.e., displaying one or more attributes of a style) pose challenges as well. Two corpora that are representative of two different attributes might well be accessible, but they might have little content overlap (e.g., datasets of texts for children and datasets of scholarly papers), which complicates the learning of content preservation (Romanov et al., 2019). Hence, methods devised on them are alternatives to the translation-like learning, and they can be broadly grouped into two classes.

The first type of approach applies an **explicit style-to-content disentanglement** with styles that realize in language through particular markers. Regarding the style of formality, for instance, expressions like "*could you please*" or "*kindly*" are more indicative of a formal tone than an informal one. This observation underlies studies that target the text transformation on style-bearing words (Li et al., 2018; Madaan et al., 2020; Wen et al., 2020; Xu et al., 2018; Sudhakar et al., 2019; Lee, 2020; Malmi et al., 2020). As a preliminary step, they identify the portion of a sentence responsible for its style: removing that would produce a style-agnostic representation for the input. Next, they generate the paraphrases by focusing the re-writing edits on such portions, namely, by replacing those words with markers of a different attribute.

Techniques for explicit word replacement are relatively transparent, but they cannot be extended to all styles, many of which are too nuanced to be reduced to keyword-level markers. Methods for **implicit style-to-content disentanglement** overcome this issue (Fu et al., 2018; Lin et al., 2020; Hu et al., 2017; Shen et al., 2017; Zhao et al., 2018; Prabhumoye et al., 2018). The idea behind this second type of approach is to strip the input attribute away at the level of the text latent representations (rather than at the word level). This usually involves an encoder producing the latent representation of the input devoid of any style-related information, and a decoder generating the text guided by training losses controlling for the output attribute and content.

According to some studies, both the explicit and the implicit separation of style and content can be sidestepped (Lample et al., 2019). Arguing that this disentanglement is not only difficult to achieve (given the fuzzy boundary between the two) but also superfluous, they have abandoned its venture. They have opted for designing training losses that encourage the presence of the three criteria in the output, for instance in a reinforcement learning setup (Luo et al., 2019; Wu et al., 2019a, i.a.), and they have devised methods that condition text generation on pretrained supervised representations of the target attributes (Lample et al., 2019; Smith et al., 2019).

3.2 Emotions among Other Styles

The styles of style transfer are the most disparate, ranging from formality (Rao and Tetreault, 2018) to the style of song lyrics (Lee et al., 2019), from diachronic language varieties (Romanov et al., 2019) to the writing profile of specific novelists (Krishna et al., 2020). Emotions have found little space among those. They have sparked research on conditioned text generation (Zhou and Wang, 2018; Song et al., 2019; Huang et al., 2018, i.a.), but the multifaceted ways in which they realize in language – e.g., explicit mentions, implicit pointers, descriptions of salient events – seem to place this phenomenon at the interface between *what* is said and *how* that is done. For this reason, their transfer turns out particularly challenging.

A few studies however exist (e.g., Dryjański et al., 2018). To understand how they fit in the field, it is useful to go beyond a review of the methods devised to fulfill the three style transfer criteria. One can look at how the styles of style transfer relate to each other to clarify the features they share, and frame emotion as a style among others.

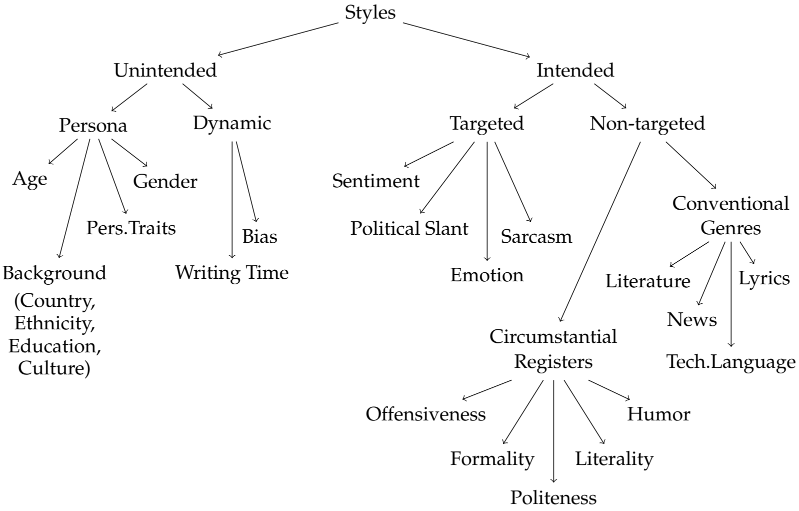


Figure 2.5: A hierarchy interconnecting the styles (outer edges) found in the style transfer literature. Pers.Traits: Personality Traits. Tech.Language: Technical Language.

3.2.1 A Hierarchy of Styles

Researchers in style transfer rely on a conceptual distinction between meaning and form (e.g., De Saussure, 1959), where the latter is a dimension of sociolinguistic variation that manifests itself in syntactic and lexical patterns, that can be correlated with independent variables and that, according to Bell (1984), we shift in order to fit an audience. Bell's characterization emphasizes the intentionality of language variations, accounting only for the styles ingrained in texts out of purpose. Yet, many others emerge as a fingerprint of the authors' identities, for instance from unambiguous markers of people's personalities and internal states (Brennan et al., 2012).

These insights can be followed to organize the publications in the field[15] from 2008 to 2021, and obtain a cohesive (and more theoretical than technical) view on it. In Figure 2.5, styles are separated into the *unintended* and *intended* families, corresponding to accidental and voluntary categories of styles. They further branch out into other groups,

[15] Troiano et al. (2022b) provide the full list of publications on which this discussion is based.

down to the most external leaves, which represent the individual styles found in style transfer. This is only a possible way of placing such leaves (e.g., writers could voluntarily foreground or silence their local dialect, that I consider *unintended*). The arrangement here is guided by the data used in the respective articles and the assumptions these put forward.

Unintended Styles. The left branch of the hierarchy includes styles that do not manifest via communicative strategies set in place by the writers. They are indicators of other variables, like one's mental disposition, biological, and social status, revealed by stylometric cues in a text. For instance, people's vocabulary becomes more positively connotated in older ages (Pennebaker and Stone, 2003); sub-cultures express themselves with a specific slang (Bucholtz, 2006). Under the assumption that writers leave a trace of their personal data, without attempting to mask their casual language use (Brennan et al., 2012), works focus on stable traits defining systematic differences between writers, or on short-lived qualities that emerge within a subject in response to situations (Beckmann and Wood, 2017). The figure refers to these between-persons and within-person language variations as *persona* and *dynamic states* respectively. *Persona* includes biographic attributes, like *personality traits*, *gender and age*. *Dynamic states* include qualities that characterize writers only in particular contexts, such as *bias* (in the sense of a subjectivity emerging when personal assessments should be obfuscated as much as possible) and the *writing time*. With similar styles, the transfer shifts the attribute young to old, Bachelor to PhD, Caucasian to Asian (Kang and Hovy, 2021), among others.

Intended Styles. The second main category of styles is *intended*, as it covers deliberate linguistic choices with which authors deploy their communicative purpose or adapt to the environment. Among these are styles used to express how one feels about the topic of discussion: a speaker/writer can have a positive sentiment *on* a certain matter, be angry or sad *at* it, be sarcastic *about* it, etc. Of this type are styles *targeted* towards a topic, like *sarcasm* and *emotion*, which pertain to the language of evaluations.

The *non-targeted* ones are non-evaluative (or non-aspect-based) styles, closer to what theoretical work calls "registers" and "genres" (e.g., following Lee, 2001). The culturally-recognized categories to which we assign texts are *conventional genres*. They can be thought of as *conventional* writing styles because they are socially coded, tailored to the ideal addressee of the message rather than an actual one, and are typically employed in mass communication (i.e. novels, poems, technical manuals, and all such categories that group texts based on intended audience or purpose of production). Their transfer includes objectives like the versification of prose, the satirization of novels, or the simplification of technical manuals. Linguistic patterns that arise in particular situations form a separate group in the hierarchy. They are *circumstantial registers*, linguistic varieties solicited by an interpersonal context, each of which is functional for immediate use. Like genres, *circumstantial registers* have distinctive lexico-grammatical patterns – e.g., the distribution of pronouns and nouns differs between a casual conversation and an official report (Biber and Conrad, 2009), but they convey a general attitude of the writers, a tone in which they talk or a social posture. An example is formality, that speakers increase if they perceive their interlocutor as socially superior (Vanecek and Dressler, 1975).

3.2.2 Where do Emotions Fit?

In Figure 2.5, I place studies dealing with emotions among the *intended* set of styles (corresponding to the leaf *emotion*), because (1) they use data consciously produced by writers around emotion-bearing impressions, (2) emotions are targeted in the sense that they are intentional or relational, and (3) they stem from evaluations. However, their placement is open to debate. Emotions have some features close to the *unintended* side of the hierarchy, for people are not necessarily aware that emotions seep out of their written productions, nor do they purposefully experience them (emotions are reactions to salient events (Scherer, 2005)). Being short-lived experiences, they could be among the *dynamic states*. Lastly, the open verbalization of an emotion might depend on the situation in which an utterance is produced (e.g., Who are my interlocutors?), and might be limited to a mention of the personal state, with no reference to its object, more in line with the *non-targeted* subset of styles.

What matters is that all these commonalities point out that emotions can be conceptualized along the same coordinates used for other leaves in the hierarchy: at first glance, the states of anger, disgust, joy and so on, constitute alternative values that are to “emotion” like formal and informal are to “formality”. Similar to other linguistic properties treated as stylistic attributes, one could expect they manifest through specific linguistic patterns. Therefore, we will assume that emotion is a style until proven otherwise.

3.3 Transferring Targeted Styles

The presence of writers in language becomes particularly evident when they assess a topic of discourse. They applaud, disapprove, convey values. Communications of this type, which pervade social media, have provided fertile ground for the growth and success of opinion mining in NLP. Opinion mining is concerned with the computational processing of stances and emotions targeted towards entities, events, and their properties (Hu and Liu, 2006). The same sort of information is the bulk of study for the *targeted* group in the hierarchy of Figure 2.5.

Dealing with evaluations makes the transfer of *targeted* styles extremely troublesome. To appreciate what is at stake here, let us take an example that explicitly mentions an emotion, "*I'm happy for you*". A style transfer task might generate a paraphrase that expresses another state, for instance sadness, by changing the emotion word into, e.g., "*sad*". Would such a modification change the stylistic attribute and preserve the meaning of the input? This question urges attention: to date, it is unclear whether this research line can aim at satisfying the three criteria to perform style transfer. Works in the field have not provided an answer, nor have other studies in NLP offered key insights. As a matter of fact, some of the styles at hand are cognitive concepts whose realization in text is yet to be fully understood (i.e., Are they content or style, or both?). The problem arises not only with input texts containing explicit markers of style (e.g., "*happy*" for emotions). Even when attitudes are expressed less directly in a sentence (e.g., "*I managed to pass the exam*"), the issue of shifting its stylistic attribute, and only its stylistic attribute, remains. Current studies solely suggest that the transfer is effortless for some texts but not for others, and that it can occur through various strategies, not necessarily by swapping emotion words (Helbig et al., 2020).

Below is a summary of past work in emotion style transfer, covering the data, the methods and the evaluation practices that have been put into use.

3.3.1 Emotion Transfer

The transfer of emotions requires rewriting a source text such that the output conveys the same message and a new emotional nuance. Source and target attribute labels can be borrowed from various traditions in psychology. While classification-based emotion analysis often uses dimensional schemas, style transfer has only relied on discrete psychological models and has limited the mapping to emotion categories. Given a source sentence like *"Damn, I broke the ancient vase!"*, that a writer might associate to a sad or angry circumstance, a joyful counterpart could be *"Yey, I finally got rid of that old wreck."*. There are also publications that do not follow any established emotion model. Lample et al. (2019) performed the transfer between discrete feelings, i.e., relaxed and annoyed. Smith et al. (2019) preferred a richer set of labels that mix affective states and emotions. They put all of them under the umbrella term of "sentiment", despite including more fine-grained labels than polarity, such as the states of being annoyed, ecstatic and frustrated.

In this research panorama, Chakrabarty et al. (2021) took a special approach. Rather than concentrating on emotions per se, they considered their *appeal* as an argumentative strategy that makes texts persuasive to an audience. These authors paraphrased textual arguments with strong emotional effects (e.g., *"At this dire moment, we all need to amplify our voices in defense of free speech."*) into more "trustworthy" variants devoid of visceral appeal (e.g., *"At this crucial moment, we all need to amplify our voices in support of free speech."*), for instance without any hint to fear.

Since emotions are ubiquitous in communication, there is an unbounded number of applications where their transfer could be applied, from clinical to political contexts. In keeping with Chakrabarty et al. (2021), style transfer tools could strip emotions away from existing arguments in order to isolate their factual core; vice versa, they might enhance the persuasive power of arguments by infusing them with a specific emotion. They could give an emotional slant to learning materials in the domain of education, to stimulate learning processes (Zull, 2006). Augmenting emotions or making them explicit might also facilitate textual understanding for individuals who struggle to interpret the expression of affective states, like people with high traits of autism or alexithymia (Poquérusse et al., 2018). In commerce, they could rewrite trailers of books, movies or the presentation of any other product, with higher emotional impact. Lastly, any chatbot capable of emotion transfer may adjust the affective connotation for the same semantic gist depending on its users.

Data. Of the comparably large set of emotion corpora compiled from various domains (Bostan and Klinger, 2018), only a small subset has served style transfer. Among them are TEC, the corpus of Tweets from Mohammad (2012), and the EMPATHETICDIALOGUES dataset from Rashkin et al. (2019) used by Smith et al. (2019), which encompasses a wide range of mental states. A corpus that is not dedicated to emotions but contains them as personality-related labels is the PERSONALITY-CAPTION dataset (Shuster et al., 2019), leveraged by Li et al. (2020).

Concerning emotions and arguments, Chakrabarty et al. (2021) collected 301k textual instances from the subreddit *Change My View*, a forum dedicated to persuasive discussions. They created a parallel corpus with the help of a masked language model and a resource that labels nouns and adjectives with their connotations, including the label "Emotion Association" (Allaway and McKeown, 2021). The authors matched the words in the arguments that they collected to the entries in such external dictionary, masked those associated with fear, trust, anticipation and joy, and constrained the replacements proposed by the language model to have a different emotional association than the original one.

Methods. Being an under-explored task, emotion style transfer was tackled by Helbig et al. (2020) with a pipeline that allows transparent investigation. They devised an approach for explicit content-to-style disentanglement motivated by the shortage of parallel emotion data. The pipeline's subsequent components (1) identified textual portions to be changed, (2) found appropriate new words to perform lexical substitutions, and (3) among the alternatives paraphrases resulting from such substitutions, picked one with the best fluency, content preservation and presence of target attribute. Each step was instantiated with many strategies. For (1), a rule-based identification of words vs. a selection mechanism informed by the attention scores of an emotion classifier; for (2), retrieving new words from WordNet vs. leveraging the similarity between input embeddings and those of possible substitutes; for (3), re-ranking the outputs with different weights for the three transfer criteria. The approach of Dryjański et al. (2018), who used a neural network to perform phrase insertion, is similar to that of Helbig et al. (2020) in the idea that the changes should interest only specific portions of texts.

Also Chakrabarty et al. (2021) generated multiple styled rewritings for an input text. Their model was a fine-tuned BART which learned to generate texts on their parallel data (the artificially-created text being the input and the original argument representing the target). Generation was further controlled by inserting a special separator token as a delimiter for the words that the model needed to edit during fine-tuning. In their framework, a paraphrase with the same meaning as the input was one with the highest entailment relation to the original text.

Though not directly formulated in emotion-related terms, an effort of emotion style transfer can be found in Nangi et al. (2021). There, the produced paraphrases display a different degree of excitement than the original texts, mirroring the notion of arousal in the dimensional emotion model. The study aimed at gaining control over the strength of the transfer by integrating counterfactual logic in a generative model. Their variational auto-encoder was trained with a series of loss terms to promote disentanglement and obtain two separate embeddings for style and content. Counterfactuals came into play in the form of a generation loss. This guided the model to find a new representation for the input attribute, specifically, a representation that pushed the prediction made by a style classifier (given the style embeddings) towards the target attribute.

Evaluation. In a small-scale human evaluation, Helbig et al. (2020) defined a best-worst scaling annotation task: two judges chose the best paraphrase for a given sentence among four alternatives generated from different pipeline configurations.

Consistent with the idea of making arguments more trustworthy, Chakrabarty et al. (2021) conducted an evaluation analysis in which workers from Amazon Mechanical Turk rated arguments with respect to the presence of fear, while simultaneously taking into consideration the preservation of meaning (i.e., a trustworthy text would have been penalized if it had altered the input content).

4 Emotions and Meaning

While the equivalence between emotions and sentence-level style has breached NLP only recently, a more long-dated idea is that talking about emotions is talking about a dimension of meaning (Kamps and Marx, 2001). In the last century, the attempt to capture the semantic content of language stimulated the production of multiple theories, with the two predominant schools being compositional and lexical semantics.

For the former, meaning construction sits on the fence between syntax and semantics, in such a way that what an expression means (often intended as "what truth value it has") is systematically connected to its syntactic form (Kratzer and Heim, 1998). Montague Semantics, virtually the most influential of such systems, adheres to the compositionality principle, characterizing the meaning of a whole (expression) as a function of the meanings of individual parts (words, phrases) and their syntactic combinations (Montague, 1970). Proponents of compositional semantics (e.g., Partee, 1973; Janssen and Partee, 1997; Jacobson, 2014) focus on sentential meaning. The quest for principled descriptions of word meanings defines a venture on its own. Philosophers like Wittgenstein (1922), Tarski (1983) and Carnap (1988) intertwined lexical semantics to the notion of "reference" in a non-linguistic world – or possible worlds. Semanticists, instead, mainly focused on the systematic relations between words and word classes (Evens et al., 1980; Cruse, 1986) and on their componential features (Nida, 1979; Coseriu and Geckeler, 1974). Dictionary-based studies of emotion are the affective counterpart of this lexical current. They allocate dictionaries of words with attitudes or values, used in the automatic treatment of emotions, hinging on the idea that next to a somewhat objective meaning there is a subjective one that can be measured (Osgood et al., 1957, 1975).

Providing a comprehensive list of formalisms to tackle meaning is beyond the scope of the thesis. What matters for my studies is one particular approach to semantics that is close to the constructs of emotions in many regards, mainly because it orbits around the notion of event: event semantics. Earlier, we regarded events as the seed of any emotion episode. In this section, they represent a linguistic yardstick to understand utterances. Event semantics scrutinizes the structure of sentences in a middle ground between the lexical and compositional streams of research. It focuses on word meaning while accounting for as much sentential context as necessary to understand events. To introduce it, I will give special consideration to FrameNet. This database has not been used in emotion analysis, but I will highlight that it has evident points of junction with affective phenomena, both in its internal organization and in the task of semantic role labeling that it supports.

4.1 Event Semantics

Events are central to much linguistic theorizing based on the seminal Davidsonian work. According to Davidson (1967), events consist of "spatio-temporal things", as entities that denote certain types of verbs. Action verbs like "*died*" or "*stab*" postulate an existential quantification, and therefore involve an ontology of events (Pianesi and Varzi, 2000): "'That Caesar died' is really an existential proposition, asserting the existence of an event of a certain sort, thus resembling 'Italy has a king', which asserts the existence of a man of a certain sort" (Ramsey, 1927). This paradigm has been given new breadth by the Neo-Davidsonian turn (e.g., Higginbotham, 1985, 2000). There, the "things" that are taken as events multiply. They are not confined to action verbs, but span over predicates in general, from accomplishments to achievements, from states to processes – in Bach (1986)'s words, to any eventuality. To date, this stance is a standard for event semantics.

An important aspect of events is their relational character. They presuppose participants that serve some functions (Maienborn, 2011). Therefore, to understand events in language is to recognize what happened (typically expressed by the predicate of a clause) as well as what properties it had (its arguments), such as where, when, how the event occurred, and who was involved in it (Màrquez et al., 2008). This idea pins down the goal of event semantics, which is to capture structured characteristics of sentences, to unfold their semantic properties via lexical and syntactic ones.

A way of putting events and participants in relation is through thematic roles. Thematic roles are two-place relations between arguments and events, which support a systematic representation of functions, sets of entailments, and properties (Davis, 2019). They are also at the core of Semantic Role Labeling (SRL) in NLP, the task of automatically identifying relations between a predicate and its arguments (Gildea and Jurafsky, 2002).

Given a sentence, a semantic interpreter assigns words to the thematic roles they fill, like the subscripts in the following examples:

(1) [My friend Julia $_{\text{Agent}}$] opened [the bottle $_{\text{Theme}}$] [with a knife $_{\text{Instrument}}$].

(2) [The bottle $_{\text{Theme}}$] was opened [by an expert sommelier $_{\text{Agent}}$].

These texts illustrate how roles express a predicate–argument relation without being bound to the specific linguistic instantiation of either. The same role can be held by different syntactic and semantic elements. The argument of a verb can even occupy different syntactic positions – a phenomenon called "diathesis alternation" (Jurafsky, 2000, Chapter 5). Put another way, roles capture commonalities between parts of sentences that share no material at the surface level. They abstract the structure of events away from its verbal realization, allowing to infer, for example, that *My friend Julia* in (1) and the *expert sommelier* in (2) are both actors willingly performing an action. Parsing roles thus produces shallow meaning representations, conceptual extrapolations that are not deeply rooted in their syntactic structure. This does not mean that they provide shallow meaning information. On the contrary, roles give access to meaning irrespective of diathesis alternation phenomena (e.g., the bottle in the examples above has been opened by someone, regardless of whether this someone is a subject or a prepositional complement).

The particular set of roles that a SRL system outputs depends on the adopted theoretical model and the corresponding inventory of predicates and roles of interest.

4.2 FrameNet

One strand of formalisms to model predicate-argument structures has linked the definition of roles to the verbs, or groups of verbs, they accompany. To this tradition belong two well-established semantic lexicons: PropBank and FrameNet. In PropBank (Palmer et al., 2005) are roles like Arg0 or Arg1 defined against the backdrop of Dowty's proto-roles (1991). The numbers indicate that an argument acts more as an agent (i.e., is a PROTO–AGENT), as a patient (i.e., a PROTO–PATIENT), or as any other type of thematic role (e.g., INSTRUMENT, ATTRIBUTE). These roles are specific to the sense of individual verbs. They support the understanding that *My friend* is to *open* in (1) as the *sommelier* is to *open* in (2), because both of them are Arg0 (i.e., PROTO–AGENT) of the same verb.

Another formalism comes from the theory of frame semantics (Fillmore, 1982), in which roles are not specific to verbs only. They define a more granular and interpretable semantic object. In addition to similarities across diathesis alternations already captured in PropBank, frame semantics roles permit to find similarities among arguments of different predicates (e.g., of "*revealed*" and "*disclosed*" below):

(3) [The ministry $_{\text{Agent}}$] revealed the numbers during a press release.

(4) In a press release, numbers were disclosed by [the ministry $_{\text{Agent}}$].

This theory upholds that utterances are understood as they evoke situations (i.e., frames) of which we know the structural properties (i.e., their roles, or frame elements). A frame represents a situation fragment that serves to match a word or a group thereof to the bundle of knowledge it presupposes (Ruppenhofer et al., 2016). For instance, we comprehend that a "*revelation*" takes a piece of information to be brought to light and someone making it public, and that these components occur with a "*disclosure*" as well. Likewise, the term "*abandon*" evokes a conceptual category instantiated by many events (e.g., leaving a membership group, or metaphorically, quitting a bad habit) comprising a series of participants (e.g., the group being left, the person dropping out of it): the evoked frame, namely "abandonment", conjugates all these bits of knowledge. In this context, traditional thematic roles like AGENT and PATIENT are elements defined by frames (Fillmore and Baker, 2000). They constitute the scaffold making the frames up, and they can be core roles specific to a frame, or non-core roles shared across many of them.

To organize the frames evoked by events together with the participants they involve, the Berkeley FrameNet project (Baker et al., 1998) has created FrameNet, an online lexical resource for English. FrameNet provides an inventory of predicates ("lexical units"), roles (arguments), and frames, and it documents relations between them. Some are of a linguistic nature, others are conceptual relations between classes of different events. In its latest release (FrameNet 1.7), it has over 1.2k frames and 13k lexical units.

An example for the frame ABANDONMENT contained in the database[16] is shown in Table 2.2. ABANDONMENT can be evoked by verbs (boldfaced in the examples (5), (6) and (7)), but also by other lexical units such as adjectives and nouns – i.e., predicates that are part of a coherent semantics. It has the roles of AGENT and THEME representing the "frame elements" that participate in the situation, where the former expresses the entity leaving the latter. Moreover, this frame links to the INTENTIONALLY_AFFECT via an INHERITANCE relation. That is, it inherits properties from this broader conceptual class, and can thus be considered a specific kind of INTENTIONALLY_AFFECT situation.

[16] Frame definitions at: `https://framenet.icsi.berkeley.edu/fndrupal/frameIndex`.

Frame: Abandonment	
Definition	An Agent leaves behind a Theme effectively rendering it no longer within their control or of the normal security as one's property.
Lexical Units	abandon.v, abandoned.a, abandonment.n, forget.v, leave.v
Elements	Agent, Theme
Example Sentences	(5) Perhaps [he $_{\text{Agent}}$] **left** [the key $_{\text{Theme}}$] in the ignition. (6) [She $_{\text{Agent}}$] **left** [her old ways $_{\text{Theme}}$] behind. (7) **Abandonment** [of a child $_{\text{Theme}}$] is considered to be a serious crime in many jurisdictions.

Table 2.2: Example of a FrameNet frame. In the three example sentences, boldfaced words are frame-evoking predicates, bracketed words are arguments.

The examples (5), (6), and (7) in Table 2.2 illustrate the sentence-level annotations that FrameNet provides, useful for semantic role labeling.

4.3 Methods for Semantic Role Labeling

SRL systems have been proposed with probabilistic (Sutton and McCallum, 2005; Chen et al., 2008) and machine learning (Zhou and Xu, 2015; Li et al., 2021) solutions. Most works train systems in a supervised fashion, which requires role-annotated data. Typically, the end goal comprises various subtasks (Cai et al., 2018). One is the predicate identification, another is the detection of the boundaries of the verb predicate for the subtask of argument identification, where the argument can be a contiguous sequence of words, but not necessarily. While this objective is more related to syntax, argument classification has to do with semantics: roles are labeled with a function that assigns probabilities to the roles for the argument candidates. These subtasks are typically followed by a joint scoring that re-assesses the local predictions and produce the final argument structure for the predicate, for instance by leveraging dependencies between many of its arguments or by verifying that certain linguistic and structural constraints are respected (Punyakanok et al., 2008; Toutanova et al., 2008). There are exceptions to this framework. Studies can address the steps in sequence using the output of a subtask as the input to the other, or skip some steps (Màrquez et al., 2008).

Early approaches relied on feature engineering, such as constituent-based and dependency-based features (Pradhan et al., 2005; Li et al., 2010), but to boost the systems' efficiency and generalization to out-of-domain data, research has shifted to neural networks. The pioneering article of Collobert et al. (2011) presented an end-to-end approach that models the full context of a word with a convolutional neural network, and classified semantic roles with a conditional random field. Following that work, recent approaches to SRL are made in an end-to-end fashion. Cai et al. (2018), for instance, accomplish all subtasks with a single graph-based model, and Swayamdipta et al. (2016), Fernández-González and Gómez-Rodríguez (2020), and Fei et al. (2021) cast a transition-based framework that incrementally finds predicates and arguments.

What splits the field is the importance of syntactic information. Many studies adopt syntactic parsing to model long-range dependencies between predicate–argument constituents, proving that syntactic supervision is beneficial to the task (Täckström et al., 2015; Roth and Lapata, 2016; Kasai et al., 2019; Cai and Lapata, 2019). Others stress the idea that syntactic dependency is a demanding goal per se, which can be removed from SRL because deep models implicitly capture that information. A number of models have been implemented that excel in role labeling while reducing the use of syntactic information (He et al., 2017; Marcheggiani et al., 2017).

Part of the literature banks on the frame semantics paradigm (Erk and Padó, 2006; Yang and Mitchell, 2017; Das and Smith, 2011). A number of systems for FrameNet-based SRL have been made available as off-the-shelf tools. Among them are the role labeler that leverages sentence and discourse context by Roth and Lapata (2015), the probabilistic models of Das et al. (2010) which use latent variables of lexical-semantic features to facilitate frame predictions for unknown predicates, and the interpreter of Swayamdipta et al. (2017) that detects FrameNet frames and frame elements.

4.4 Emotions and Frames

The development of FrameNet took place around several predefined domains. Emotions were one of them. In a thorough emotion-oriented analysis of this database, Ruppenhofer (2018) discusses the criteria that guided the placement of lexical units under specific frames, like the constraint that lexical units in a frame should accept the same types and number of syntactic dependents, or that the frames EXPERIENCER_SUBJ and EXPERIENCER_OBJ should include predicates where the emotion experiencer has the role of subject and object, respectively. Such criteria allow words indicating different emotions to fall within the same group of predicates. Conversely, words with the same lexical root can be part of different frames. For example, the verbs "*please*" and "*anger*" are part of EXPERIENCE_OBJ, in which the emotion experiencer takes on the role of the object, but the adjective "*angry*" is not. There is in this sense a key difference between a dictionary-based and a frame semantics account of emotions. Namely, the organization of emotions supported by the latter reflects the similarities between their linguistic realizations, more than their glossary characterization.

Both emotions and frames stem from events and can be understood against their knowledge. Yet, frame-based semantic parsers have never been explored to study emotions. A few research lines in emotion analysis only come near the study of semantic roles. Structured emotion predictions are one of them. Based on the understanding of emotion events in terms of semantic roles, they are in agreement with semantic parsing. For the latter, understanding how events are realized in language is to recognize what happened and what properties it had. Similarly, emotion role labeling is interested in who feels what and why (Kim and Klinger, 2018). Its task is to automatically assign textual spans to roles like experiencers and stimuli (Mohammad et al., 2014; Oberländer and Klinger, 2020; Oberländer et al., 2020). While disregarding frames, it identifies the portions of texts corresponding to emotion causes, emotion holders, and eventually, the entity towards whom the emotion is directed (e.g., "*I am mad at you*"). Frames and emotion semantic roles are brought together by the work of Ghazi et al. (2015), who created a corpus in which 820 FrameNet sentences are annotated with emotions and the semantic role of stimulus. However, the corpus was built using texts that contain lexical units only coming from the frame EMOTION_DIRECTED, and it neglects different semantic roles.

In summary, the last two sections touched upon two salient points of the dissertation. One is the task of emotion style transfer, which I will leverage next to text classification. I described its goal and desiderata, typically achieved by considering the transfer as a problem of translation between attributes, or by designing special training functions, or by assuming that style lurks in certain portions of texts and is to be transformed with localized edits. I further placed the existing styles of style transfer in a hierarchy that puts them in relation. Among those, **emotions pose special challenges** in the simultaneous achievement of content preservation, transfer accuracy and naturalness within the generated paraphrases.

Lastly, I took up an idea intrinsic to dictionary-based studies in emotion analysis, according to which emotions in language are affective meanings. I covered an approach to semantics focused on events, and for that, close in spirit to the study of emotions. I described how **FrameNet provides an abstract semantic system** that characterizes both predicates and arguments, already used for the automatic task of frame-based semantic role labeling, but **never-before applied in computational emotion analysis**.

Chapter 3

Emotion Detection from Implicit Expressions

This chapter faces a foundational question: how well humans recognize emotions from texts that mention none overtly. Clear examples are descriptions of emotion-eliciting events, since emotions are intrinsic to them. The link between emotions and events is studied by appraisal theories in psychology via *evaluations* of event properties (cf. Page 21). They argue that an emoter's appraisals guide her reaction to certain facts (e.g., novel situations with uncertain outcome can induce fear), and are then inferred by the observers to grasp the emoter's affective state. I bring the idea into the domain of language, tackling the annotators' ability to judge emotions and appraisals from implicit emotion expressions.

My analysis elaborates on two annotation campaigns, each made in two phases: first, people describe events in which they felt an emotion; next, readers reconstruct emotions and appraisals from such texts. In the resulting crowd-enVENT corpus, appraisal variables turn out to reveal otherwise unintelligible patterns of judgments (e.g., the evaluations behind the coders' emotion choices). Even without this layer of annotation, collecting the writers' ground truth provides a favorable condition to reflect on the inter-subjective interpretation of emotions. The resources deISEAR and enISEAR allow to do so in a multilingual setup.

Highlights

Appraisals reveal patterns of judgments, useful for a qualitative understanding of (dis)agreements.

Writers' emotion labels allow to investigate the emotion recognition ability of readers.

1 The Need for New Data

Computational emotion analysis counts on vast corpora labeled by humans in English, and when creating new resources, it is armed with actions to heighten inter-annotator agreement. Having coders indicate the emotion they are *evoked* by a text (not one they *associate* it with) is an example (Mohammad and Turney, 2013). Another practice refrains from collecting the coders' own reactions to the texts, but asks them to imagine and report on the emotion likely felt by the writers (Buechel and Hahn, 2017b).

The shortcomings surrounding data in the field are many nevertheless, and they propagate from the very beginning of the workflow (i.e., data creation) through the subsequent steps of understanding agreement and learning automatic models. There are at least three issues that concern the way in which annotations are collected, the type of data under consideration, and the theoretical models of reference. These open as many investigation tracks searching for different annotation procedures, corpora and theoretical models than those currently in use.

1.1 Issues in Annotation Procedures

About the ability to decode emotions, psychology has long motivated comparisons among the observers' judgments, and between them and the actual experiences of the emoters (cf. Chapter 2, Page 29). In computational emotion analysis instead, the question of how well annotators cope with emotion recognition is uncharted water. Researchers usually resort to readers as annotators, and the pre-existing texts they use have no ground truth information, i.e., labels released by the texts' authors indicating the emotion that their words encode. That is problematic. It means that the consistency of the readers, or how much their judgments converge, can be measured; their accuracy, or whether they are correct, cannot.

The situation is only partially solved by sourcing data on social media. Hashtags and emoticons take the role of ground truth proxies (Mohammad, 2012; Qadir and Riloff, 2014; Felbo et al., 2017), but little evidence could support that pre-existing texts were spurred by a particular affective state or were meant to communicate one. Plus, the tags' match with the emotion tone of the utterances is questionable – they could actually be illocutionary indicators of sarcasm, and alter an utterance affective attitude (Derks et al., 2008).

To complicate the picture, the coders' consistency can shrink if the texts (hint to but) do not openly point out an emotion (e.g., *"I finally found a friend"*, *"She heard a sinister sound"*). Bostan et al. (2020), for instance, obtained Cohen's κ (1960) as low as .09 in the annotation of news headlines that mention factual circumstances.

1.2 Issues in Existing Corpora

Depictions of factual circumstances in language are the verbal counterpart of an emotion pillar, i.e., events. As such, they permit to infer emotion reactions, allowing to do so also when they contain no affective marker (e.g., *"I ~~finally~~ found a friend"*). Corpora of implicitly affective texts are thus desirable from a computational standpoint. They would give ground to wide-coverage automatic recognition, with systems capable of sensing event-related emotions. Works that acknowledge the event–emotion link exist. They are a special case of semantic role labeling, for which systems detect emotion stimuli or other structural aspects of emotions mentioned in text (Kim and Klinger, 2018; Mohammad et al., 2014; Xia and Ding, 2019). However, the specific focus on implicit expressions characterizes only a handful of tasks and resources (Balahur et al., 2011; Klinger et al., 2018), and studies interested in the emotions of factive statements, of the sort of Bostan et al. (2020), are even sparser.

Other than a focus on these types of expressions, the field lacks cross-lingual comparability (Navas Alejo et al., 2020), which hampers attempts of data-driven multilingual modeling, and the development of predictive models beyond English. English, with its large and diverse datasets, is counterbalanced by a majority of low-resource languages as far as corpora go – the case of lexicons is different (Buechel et al., 2020). For instance, a few dictionaries have been created in German, like BAWL–R, a list of words rated with arousal, valence and imageability features (Vo et al., 2009), and DENN-BAWL, that extends the former to scores of emotion intensities (Briesemeister et al., 2011). Sentence-wise emotion annotations started to be available only recently (Lamprinidis et al., 2021). Still, German resources are rarer than English ones.

1.3 Issues in the Use of Theoretical Models

The paucity of event-centic corpora in computational emotion analysis complements scarce attention to psychological models where events are pivotal for emotion understanding. This claim has already been maintained in the Background discussion (Chapter 2): annotating discrete emotion categories and dimensions of affect overlooks that emotions are reactions to events evaluated by people. Extending on that argument, one can see how NLP might capitalize on the functional components that unpack the event–emotion progression: all of them manifest in language, for example with descriptions of the verbal (e.g., "*oh, wow*") or motor responses to a situation (e.g., "*I felt paralized!*") (De Bruyne et al., 2021; Casel et al., 2021). Appraisals deserve however special consideration for emotion recognition, as they are a potentially crucial tool to extrapolate affective imports from text.

Beyond text, these cognitive dimensions not only take place during emotion production in the emoters, but also help emotion decoding for the observers. For emotion production, scientists like Smith and Ellsworth (1985) and Scherer (2005) qualify appraisals with detailed event evaluation criteria, whose different combinations correspond to different emotions (e.g., the perception of an event as unpleasant and hampering one's goals could elicit anger; an event high in unpleasantness and unexpectedness, on which one has no control, could induce fear). Flipping the viewpoint to consider emotion recognition, people's empathy and the ability to assume the affective perspective of others is guided by their assessment of whether an event might have been important, threatening, or convenient for those who lived through it (Omdahl, 1995). Observers can infer how an event has been experienced by an individual, and in consequence figure out what she feels, for instance by looking at her facial muscles configuration (Mortillaro et al., 2012). Works in psychology do not particularly target verbal data, but their idea can be extended to language, where words, similar to facial movements, invite the readers to conduct an evaluation, and to picture how the speaker appraised the topic under discussion to recognize her emotion.

The importance of evaluations for emotion analysis can be fully appreciated with factual statements. The affective understanding of these expressions relies on the readers' background and experience, engaging their knowledge about event participants, typical responses, possible outcomes, and world relations. It is thanks to an (extralinguistic) assessment that texts like "*the tyrant passed away*" and "*my dog passed away*" can be associated with an emotion, and specifically, with different emotions. The two sentences describe semantically similar situations (i.e., death), but their subjects can change the comprehension of how the writer was affected in either case. For example, the first text could be charged of relief and the other of sadness, and that would be only one of the possible interpretations: individuals can converge on the semantics of the texts because word meanings establish a common ground of understanding, while appraisals, hinging on people's stances and episodic knowledge, are not necessarily shared.

Therefore, appraisals expose what makes emotions difficult to agree upon. Having access to information about these event evaluations could help annotation studies explain why linguistically similar texts convey opposite emotions (e.g., "*the tyrant passed away*" and "*my dog passed away*" have different significance to one's goals) or why annotators assign incongruent labels to the same text (e.g., they appraise the described event differently).

2 Emotions and Appraisals in a Communication Framework

The open issues about annotation design, types of expressions in the corpora, and used theoretical models represent three separate gaps for computational linguistics, but in a way or another, they all prevent to grasp how well readers recognize emotions – consequently, how classification models can be expected to perform on the data that said readers annotate. Connecting the dots: emotion inferences from factual statements are extremely subjective and lead to low inter-annotator consistency; such texts are not annotated following theoretical models of evaluations that underly emotions (and that could justify the discrete labels annotators choose), with the result that low consistency has to be accepted as such and cannot be made sense of; they also lack ground truth annotations (both of discrete emotions and of the evaluations that caused them) and therefore prevent researchers in computational emotion analysis from establishing the readers' accuracy. This is the knot I attempt to untangle here.

Goal of the Studies. This chapter reports the results of two data collection endeavors that contribute to understanding the annotators' emotion recognition ability while tackling the need for (1) texts enriched with ground truth labels, and (2) determining the limits and possibilities of emotion annotation on implicit expressions that refer to events, via discrete and appraisal models of emotions.

First, I will describe how we crowdsourced crowd-enVENT, a corpus of 6600 emotion-inducing event descriptions produced by English native speakers, annotated with emotions, event evaluations (using 21 appraisals), stable properties of the texts' authors (e.g., demographics, personality traits) and contingent information concerning their state at the moment of taking our study (i.e., their current emotion). Part of crowd-enVENT is then annotated by other coders, tasked to read the descriptions and to infer how the authors originally appraised the events in question and what emotion they felt.

Next, I will introduce deISEAR, a novel collection of emotion-inducing events described by German native speakers, which deals with the problem of cross-lingual data shortage. deISEAR is a simplified version of crowd-enVENT (e.g., with no appraisal label and only 1001 data points). With the same procedure, we collected other 1001 event descriptions from English native speakers: enISEAR bridges deISEAR (with which it shares the experimental design) and crowd-enVENT (with which it shares the language). Like the latter, deISEAR and enISEAR are labeled from both writers and readers.

The chapter poses three research questions. (RQ1) How accurately do people infer emotions from event-centered texts? I address this question in both annotation studies by comparing judgments collected from external coders and text writers. The other two questions only regard the first, appraisal-based study. (RQ2) To what extent can readers infer appraisals? (RQ3) Do appraisal judgments help explain qualitative differences in emotion assignments? I scrutinize the idea that event evaluations are useful constructs for emotion analysis by investigating how coders tag the texts with appraisal dimensions.

My discussion will start by elaborating on the crowdsourcing-based approach that underpins all three corpora (Section 3). It will continue by going through different stages of the two corpora-construction activities, that is, data creation and analysis of annotations. Regarding data creation, Section 4 will introduce the collection and the resulting features of crowd-enVENT (Section 4.1), and deISEAR and enISEAR (Section 4.2). For data analysis, Section 5 will face the research questions with an analysis of inter-annotator agreement, separately for the two studies (crowd-enVENT in Section 5.1.1 and de(/en)ISEAR in Section 5.1.2).

3 Crowdsourcing-based Corpus Creation

Since we ask if (and to what extent) emotions and appraisals can be interpreted from implicit emotion expressions, our studies relate to research lines in sentiment analysis aimed at recognizing "people's opinions, sentiments, evaluations, appraisals, attitudes, [...] towards entities such as products, [...] issues, events, topics, and their attributes" (Liu, 2012). The literature on implicit expressions of polarized evaluations targets specific types of opinions, e.g., those that come forth in business news (Jacobs and Hoste, 2021, 2022) and in meeting discussions (Wilson, 2008). Much of it has the goal to understand if texts contain evaluations (Toprak et al., 2010), or how their polarity can be traced back to specific linguistic cues, like negations and diminishers (Musat and Trausan-Matu, 2010), indirectly valenced noun phrases (Zhang and Liu, 2011b) and their combination with verbs and quantifiers (Zhang and Liu, 2011a). By contrast, we do not restrict ourselves to any type of event; most importantly, we relate evaluations to people's background knowledge with a theoretically-motivated taxonomy of appraisals to make the type of evaluations behind emotion experiences and judgments transparent.

Also evaluations of events that elicit emotions have been leveraged before, but only by a handful of studies. Works like Shaikh et al. (2009), Balahur et al. (2011), and Balahur et al. (2012) proposed approaches to make emotion categorization decisions based on appraisal constructs. Yet, they did not examine the suitability of the underpinning theories for emotion annotations, which is a problem we tackle (though indirectly) by asking RQ2 and RQ3, and which poses a major challenge: there is no available corpus that contains annotations of our concern (i.e., provided by first-hand event experiencers).

Actually, a corpus to study emotions and appraisals from event descriptions is ready to use. Its development was propelled by the "International Survey on Emotion Antecedents and Reactions" (Scherer and Wallbott, 1994), conducted by a group of psychologists who collected emotion data in the form of self-reports. They administered an anonymous questionnaire to 3000 participants all over the world, tasked to recall an emotion episode that they associated with one of seven basic emotions, and to recall both their evaluation of the stimulus and their reaction to it. The aim of the survey was to probe that emotions are invariant over cultures, and are characterized by patterns of bodily and behavioral changes (e.g., change in breathing, felt temperature, speech behaviors).

This focus on private perspectives on events sets ISEAR apart from other corpora to study textual emotions. To address both RQ1 and RQ2, we could re-annotate it with the help of readers. However, its texts are unpractical, because they were produced by a combination of native and non-native speakers who only consisted of college students (for the final dataset, all the reports were translated to English). Therefore, we collect data from scratch. We build emotion and appraisal-based corpora of self-reports close to ISEAR, but via crowdsourcing and with a design comprising two phases that reproduce a simplified framework of real-life communication.

3.1 A Simplified Communication Framework

The annotation issues seen earlier can be linked to a generic dismissal of prior-to-text emotions. Chapter 2 emphasized that computational emotion analysis taps on psychology pragmatically, with approaches that do not account for what emotions are (i.e., reactions to evaluated events). In fact, studies fall short in reflecting *the situations in which emotions are communicated*. In everyday circumstances, the ground truth is available (the emotion experiencers are there), emotions emerge via appraisals, and again thanks to appraisals (among other things) they are recognized (Scherer and Grandjean, 2008; Laukka et al., 2010).

It is thus reasonable to try to give a more "naturalistic" account of emotions in text, by linking a model of emotions to a model of communication. A useful starting point is the parallel between emotion production and recognition in Figure 3.1, which shows how Mortillaro et al. (2012) compared these two sides of an emotion episode based on appraisal models. The figure also visualizes how discrete and appraisal-based frameworks differ in the two respects. Discrete models (upper quadrants) assume that a generic affective program produces an emotion, on which appraisals have no explicit or immediately causal relation. The corresponding recognition task consists in predicting an

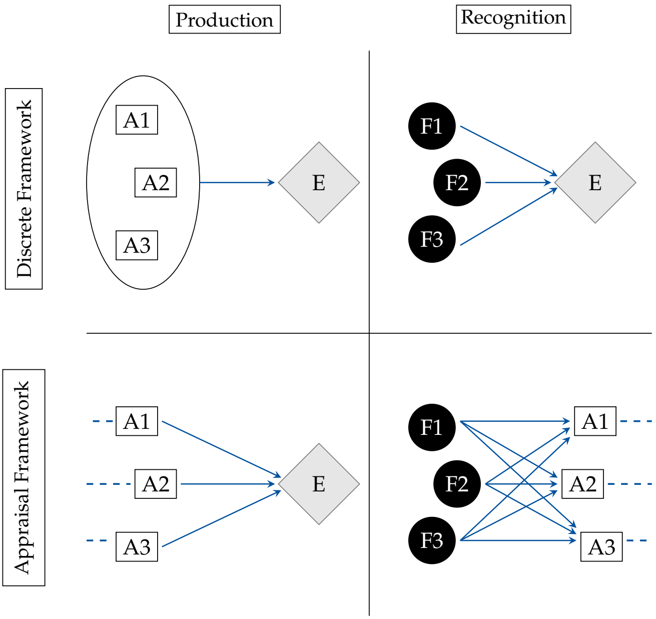


Figure 3.1: Emotion production and recognition: difference between discrete and appraisal frameworks, adapted from Mortillaro et al. (2012). Black circles are emotion expressions (e.g., facial configurations, indicated as "F"), white rectangles symbolize appraisal variables (e.g., *goal relevance*, *pleasantness*, indicated as "A"), grey diamonds ("E") are emotions (e.g., anger, disgust).

emotion label – once more, regardless of appraisals. By contrast, in the production side of the appraisal framework (bottom quadrants), certain event evaluations lead to experience certain emotions; then, observers detect the visible consequences of appraisals (e.g., blushing, facial movements) and infer an emotion by inferring such appraisals (hence the dashed lines connect the two sides of the figure).

Let us adapt it to a simplified communication framework that comprises an environment and two agents, i.e., a message source and a message receiver. When an event happens within the environment, an agent is stimulated an emotion together with multiple changes. Among them is a verbal expression. The agent's emotion state is encoded in the expression, in the sense that this can refer to any point of the emotion production process (e.g., the event, the way it was evaluated, the ensuing emotion, all sorts of corresponding symptoms). Having observed the message, the other agent works her way toward the correct understanding of the state of the source. This process corresponds to the left and right portions of Figure 3.1, where the message would be a black circle, i.e., a variable from which an emotion or various appraisals and an emotion can be inferred, depending on the psychological model one focuses on (either the top or bottom parts of the figure).

We propose a data collection procedure sensitive to these emotion production–recognition dynamics, though considering texts in isolation and not in a dialogic context. Our approach toward appraisals and discrete models is not as stringent as for theorists in psychology. Our goal is not to distinguish the two models. We use the top vs. bottom distinction in the figure just to operate within a framework or another, and to define the salient variables in each of them (i.e., only emotions, or emotions and appraisals).

3.2 Two-Phase Experimental Design

Our experimental design follows this communication framework. In an experimental setting, it approximates the natural flow of information from emotion production to recognition (some differences are discussed later), in which an agent–source (the writer) and an agent–receiver (the reader) have different and well-defined goals. The source sends a verbal message (i.e., an implicit emotion expression), and the receiver has to correctly recover the state that the message encodes. Accordingly, data collection is divided into two consecutive parts: a first phase for generating the data and a second to validate it. These are represented in Figure 3.2.

Ensuring that the agent–source writes a message under a specific affective state is challenging (e.g., it would require us to induce that state). Hence, for Phase 1, participants recollect personal events that elicited an emotion in them ((a) in the figure). They describe such events and annotate the resulting texts ((b) in the figure) with information that concerns both emotions and appraisals in the first study, and only emotions in the second. Taking the latter as an example, the task of Phase 1 can be thought of as a sampling from $p(\text{description} \mid \text{emotion})$, obtaining likely utterances for a given emotion. A difference with communication in real life is that no event spurs a mental state here; the message senders verbalize one that happened in the past and they have no direct interlocutor.

After the emotion encoding step, Phase 2 focuses on the opposite

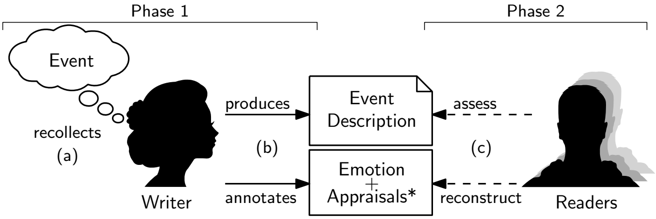


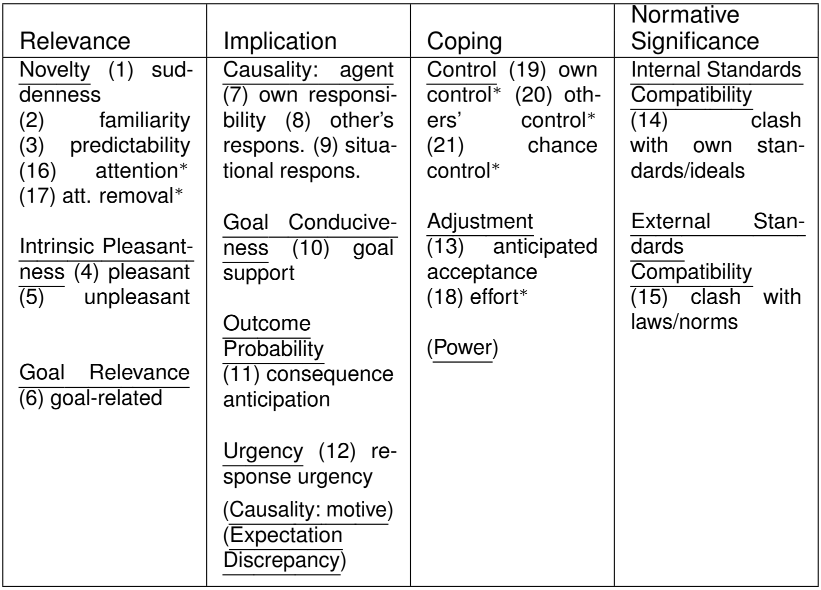
Figure 3.2: Design overview of the two corpus creation studies. The asterisk indicates that appraisals are included only in one study (for crowd-enVENT).

task, i.e., emotion decoding. Absent from the collection design of ISEAR, this stage brings validators to assess the events produced in Phase 1. They reconstruct the original emotion and, only for the first study, the original appraisals ((c) in the figure). Taking emotions as examples again, the task estimates $\arg\max_{E \in \text{emotions}} p(\text{E} \mid \text{description})$, evaluating the mapping between a given description and the emotion that corresponds it the most, out of a pre-defined set. As opposed to previous studies, we do not ask the coders to infer a generic emotion that can be attributed to the writers, but we invite them to reconstruct the writers' emotions at a specific moment in time, i.e., when the event happened. The participants' intuitions gathered this way are interpretable as a measure for the interpersonal validity of the descriptions, and as a point of comparison for automatic classification results.

We refer to the authors/writers (the agent–sender or source) of the event descriptions of Phase 1 also as "generators", and to the readers (the agents–receivers) of Phase 2 as "validators". Both are considered annotators of the texts.

3.3 Definition of Annotation Variables

Here we establish the theoretical outset of our questionnaires, describing how we defined the variables of interest: appraisals (Section 3.3.1), emotions (Section 3.3.2), and some supplementary variables (Section 3.3.3) used in later chapters of the thesis. These variables are all exploited to create crowd-enVENT (Section 4.1). A subset of them is then resumed in the smaller crowdsourcing activity of deISEAR and enISEAR – with slight differences also in the answer options presented to the annotators, as detailed in Section 4.2. A transparent comparison between the two studies is in Appendix A, Section 1.1.



Relevance	Implication	Coping	Normative Significance
Novelty (1) suddenness (2) familiarity (3) predictability (16) attention* (17) att. removal* Intrinsic Pleasantness (4) pleasant (5) unpleasant Goal Relevance (6) goal-related	Causality: agent (7) own responsibility (8) other's respons. (9) situational respons. Goal Conduciveness (10) goal support Outcome Probability (11) consequence anticipation Urgency (12) response urgency (Causality: motive) (Expectation Discrepancy)	Control (19) own control* (20) others' control* (21) chance control* Adjustment (13) anticipated acceptance (18) effort* (Power)	Internal Standards Compatibility (14) clash with own standards/ideals External Standards Compatibility (15) clash with laws/norms

Figure 3.3: Appraisal objectives (top boxes) with their relative checks (underlined) and the appraisal dimensions investigated in our work (numbered). Checks in parenthesis have been proposed by Scherer and Wallbott (1997) but are not included in our study. Items marked with an asterisk come from Smith and Ellsworth (1985).

3.3.1 Appraisals

We adopt the schema of Sander et al. (2005), Scherer et al. (2010) and Scherer and Fontaine (2013), who group appraisals into the four categories seen earlier in Figure 2.3 (Page 24), which represent specific evaluation objectives. There is a first assessment aimed at weighing the relevance of an event, followed by an estimate of its consequences, and of the experiencer's own capability to cope with them; last comes the assessment of the degree to which the event diverges from personal and social values.

Each objective is instantiated by multiple evaluation checks, and each check can be broken down into one or many appraisal dimensions. Namely, (1) *suddenness*, (2) *familiarity*, (3) *predictability*, (4) *pleasantness*, (5) *unpleasantness*, (6) *goal-relatedness*, (7) *own responsibility*, (8) *others' responsibility*, (9) *situational responsibility*, (10) *goal support*, (11) *conse-*

quence anticipation, (12) *urgency of response,* (13) *anticipated acceptance of consequences,* (14) *clash with one's standards and ideals,* (15) *violation of norms or laws.* Figure 3.3 (adapted from Scherer and Fontaine (2013)) collocates all fifteen appraisals as numbered items under the corresponding checks (the underlined texts). They constitute the majority of appraisals judged by the annotators of crowd-enVENT.

These dimensions illustrate properties of events and their relation to the event experiencers. They were used by Scherer and Wallbott (1997) to create the corpus ISEAR[1], and many of them can also be found in other studies in psychology. For instance, while formulating the questions differently, Smith and Ellsworth (1985) analyze *pleasantness*, *certainty*, and *responsibility* (they merge *others'* and *situational responsibility* together). In addition, they tackle a handful of dimensions which are only implicit in Scherer and Wallbott (1997), specifically (16) *attention,* and (17) *attention removal,* two assessments that relate to the relevance and the novelty of an event, and (18) *effort,* which is the understanding that the event requires the exert of physical or mental resources, and is therefore close to the assessment of one's Coping potential (cf. Figure 3.3). Smith and Ellsworth (1985) also divide the check of Control into the more fine-grained dimensions of (19) *own control of the situation,* (20) *others' control of the situation,* (21) *chance control.*

To obtain a large coverage of dimensions, we integrate the approaches of Scherer and Wallbott (1997) and Smith and Ellsworth (1985), adding the latter six criteria to our questionnaire. We include *attention* and *attention removal* under Novelty in Figure 3.3, *effort* as part of the Adjustment check, and *own, other's* and *chance control* inside Control. However, we disregard a few dimensions from Scherer and Wallbott (1997). In Figure 3.3, they correspond to the checks "Causality: motive", "Expectation discrepancy" and "Power". As they differ minimally from other appraisals, they would complicate the task for our annotators who, as opposed to Scherer and Wallbott's participants, do not carry out the task in lab and work under minimal training.

Research in psychology also proposes best practices to collect appraisal data. Yanchus (2006) in particular casts doubt on the use of questions that annotators typically answer to report their event evaluations (e.g., "Did you think that the event was pleasant?", "Was it sudden?"). Questions might bias the respondents, allowing them to develop a theory about their behavior in retrospect. Statements instead

[1] Questionnaire: `https://www.unige.ch/cisa/files/3414/6658/8818/GAQ_English_0.pdf`.

leave people free to recall if the depicted behaviors applied or not (e.g., "The event was pleasant.", "It was sudden."). In accordance with this idea, we reformulate the questions of Scherer and Wallbott (1997) and Smith and Ellsworth (1985) as affirmations, aiming to preserve their meaning and to make them accessible for crowdworkers. They are detailed below (appraisal names in parentheses are those I will use henceforth).

Note that only some dimensions have a semantic opposite, like *pleasantness* and *unpleasantness*. To remain consistent with the questionnaires of reference, we do not enforce this pairwise symmetric structure for all of them. However, all dimensions have to be rated for a given text. The rating occurs on a 1-to-5 scale, considering how much a dimension applies to the described event (1:"not at all", 5:"extremely"). A comparison between our appraisal statements and the original questions, with the respective answer scales, is reported in Section 1.2 in Appendix A.

Novelty Check. According to Smith and Ellsworth (1985), emotions arise in an environment that requires a certain level of attention. Kin to Novelty, the evaluation of whether a stimulus is worth attending or ignoring can be considered the onset of the appraisal process. Their study treats attention as a bipolar dimension, which goes from a strong motivation to ignore the stimulus to devoting it full attention. Similarly, we define two statements[2]:

16. I had to pay attention to the situation. (*attention*)
17. I tried to shut the situation out of my mind. (*not consider*)

Stimuli that occur abruptly involve sensory-motor processing other than attention. To account for this, the check of Novelty develops along the dimensions of suddenness, familiarity and event predictability, respectively formulated as:

1. The event was sudden or abrupt. (*suddenness*)
2. The event was familiar. (*familiarity*)
3. I could have predicted the occurrence of the event. (*event predictability*)

Intrinsic Pleasantness. An emotion is an experience that feels good/bad (Clore and Ortony, 2013). This feature does not denote

[2] The numbers preceding the statements correspond to the same dimensions in Figure 3.3.

the state of the experiencer. It is intrinsic to the eliciting condition (i.e., it bears pleasure or pain):

4. The event was pleasant. (*pleasantness*)
5. The event was unpleasant. (*unpleasantness*)

Goal Relevance Check. As opposed to Intrinsic Pleasantness, this check involves a representation of the experience for the goals and the well-being of the organism (e.g., one could assess an event as threatening). We define the corresponding appraisal as:

6. I expected the event to have important consequences for me. (*goal relevance*)

Causal Attribution. Tracing a situation back to the cause that initiated it can be key to understanding its significance. The check of causal attribution ("Causality: agent" in Figure 3.3) is dedicated to spotting the agent responsible for triggering an event, be it a person or an external factor (one does not exclude the other):

7. The event was caused by my own behavior. (*own responsibility*)
8. The event was caused by somebody else's behavior. (*other responsibility*)
9. The event was caused by chance, special circumstances, or natural forces. (*situational responsibility*)

Scherer and Fontaine (2013) also include a dimension related to the causal attribution of motives ("Causality: motive" in Figure 3.3), which is similar to the current one but involves intentionality. We leave intentions underspecified, such that for 7., 8. and 9., the agents' responsibility does not imply that they purposefully triggered the event.

Goal Conduciveness Check. The check of Goal Conduciveness is dedicated to assessing whether the event will contribute to the organism's well-being:

10. I expected positive consequences for me. (*goal support*)

Goal relevance (6.) differs from this appraisal: an event might be relevant to one's goals and needs while not being compatible with them (it might actually be deemed important precisely because it hampers them).

Outcome Probability Check. Events can be distinguished based on whether their outcome can be predicted with certainty. For instance, the loss of a dear person certainly implies a future absence, while taking a written exam could develop in different ways. Our annotators recollected whether they could establish the consequences of the event, at the moment in which it happened, by reading:

11. I anticipated the consequences of the event. (*anticipated consequences*)

Scherer and Fontaine (2013) identify one more check about consequences: people picture the potential outcome of an event based on their prior experiences, and then evaluate if the actual outcome fits what they expected. We refrain from introducing appraisals for the Expectation Discrepancy Check (under "Implication", in Figure 3.3) in our repertoire. For one, it is hard to distinguish from Outcome Probability Check in a crowdsourcing setting; but mainly, such a dimension clashes with our attempt to induce the mental evocation of their state *at the time in which the event happened* (e.g., when taking an emotion-eliciting exam), and not when consecutive developments became known (e.g., when learning, later, if they passed). Brief, 11. aims at understanding if people could picture potential outcomes of the event, and not if their prediction turned out correct.

Urgency Check. One feature of events is how urgently they require a response. This depends on the extent to which they affect the organism. High priority goals compel immediate adaptive actions:

12. The event required an immediate response. (*urgency*)

Control Check. This group of evaluations concerns the ability of an agent to deal with an event, specifically to influence its development. At times, "event control" is in the hands of the experiencer (irrespective of whether they are also responsible for initiating it[3]), other times it is held by external entities, and yet others the event is dominated by factors like chance or natural forces (Smith and Ellsworth, 1985). Accordingly, we formulate the following three statements:

19. I was able to influence what was going on during the event. (*own control*)
20. Someone other than me was influencing what was going on. (*others' control*)

[3] Vice versa, one may be responsible, but not in control of the situation (e.g., "*when I forgot to set an alarm*").

21. The situation was the result of outside influences of which nobody had control. (*situational control*)

We do not focus on "Power" (under "Coping" in Figure 3.3), the assessment of whether agents can control the event at least in principle (e.g., if they possess the physical or intellectual resources to influence the situation).

Adjustment Check. Related to control is the evaluation of how well an experiencer will cope with the foreseen consequences of the event, particularly with those that cannot be changed:

13. I anticipated that I would easily live with the unavoidable consequences of the event. (*accepted consequences*)

A different dimension of Adjustment Check is motivated by Smith and Ellsworth (1985). Emotions can be differentiated on the basis of their physiological implications, similar to the notion of arousal in the dimensional models of emotion. More precisely, subjects anticipate if and how they will expend any effort in response to an event (e.g., fight or flight, do nothing). We phrase this idea as:

18. The situation required me a great deal of energy to deal with it. (*effort*)

Internal and External Standards Compatibility. The significance of an event can be weighted with respect to one's personal ideals and social codes of conduct. Two appraisals can be defined on the matter:

14. The event clashed with my standards and ideals. (*internal standards*)
15. The actions that produced the event violated laws or socially accepted norms. (*external norms*)

The first pertains to an event colliding with desirable attributes for the self, with one's imperative motives of righteous behavior. The second concerns its evaluation against the values shared in a social organization. Both guide how experiencers react to events.

3.3.2 Emotions

Our choice of emotion categories is closely related to that of appraisals, because different emotions are marked by different appraisal combinations. In the literature, such a relationship is addressed only for specific emotions. Therefore, we motivate the selection of this variable following appraisal scholars once more. We consider the emotions that one or several studies claim to be associated to the appraisals of Section 3.3.1.

We include all emotions from Scherer and Wallbott (1997) as a first nucleus. They are *anger*, *disgust*, *fear*, *guilt*, *joy*, and *sadness* (i.e., Ekman's basic set), plus *shame*. We also use *pride*, which has to do with the objectives of Relevance, Implication, Coping potential and Normative Significance (Manstead and Tetlock, 1989; Roseman, 1996, 2001; Smith and Ellsworth, 1985; Scherer et al., 2001a). The last two works further comprise *boredom*, and Roseman et al. (1990) and Roseman (1984) examine *surprise*, as well as the positive emotion of *relief*. *Trust*, an emotion present in Plutchik's wheel, is linked to the appraisal of *goal support* (Lewis, 2001) and to the check of Control (Dunn and Schweitzer, 2005).

We define our questionnaires sampling emotions from these 12 labels, to which we add a "no emotion" category (*noemotion*), because events can be appraised along our 21 dimensions even if they elicit no affective reaction. This neutral class serves as a control group to observe differences in appraisal between emotion- and non-emotion-inducing events. However, not all texts generated for this label in crowd-enVENT will turn out to depict uninfluential or unemotional events. As pointed out later, many of them recount rather dramatic circumstances that did not stir the experiencers up.

3.3.3 Other Variables

We use two other groups of variables regarding either the described circumstances, such as emotion and event properties, or the types of persona providing the judgments, that is, features of the annotators. Note that I do not analyze all variables in this chapter: part of them will be leveraged in Chapter 4, others are only reported to provide a complete description of crowd-enVENT.

Properties relative to Emotions and Events. It is reasonable to assume that the same event is appraised differently depending on its specific instantiation. For example, while standing in a queue, an emoter of boredom could feel more in control of the situation than another, depending on how long each of them persists in it, or how intensely the event affects them. Motivated by this, we consider the *duration of the event* (with the possible answers "seconds", "minutes", "hours", "days", and "weeks"), the *duration of the emotion* ("seconds", "minutes", “hours”, “days”, and “weeks”[4]), the *intensity* of the experience (to be rated on a 1 to 5 scale ranging from “not at all” to “extremely”).

Properties of Annotators. We collect demographic data with the rationale that self-perceived belonging to a socio-cultural group can determine particular mental associations with an event. We request participants to disclose their *gender* (“male”, “female”, “gender variant/non conforming”, and “prefer not to answer”) and *ethnicity* (either “Australian/New Zealander”, “North Asian”, “South Asian”, “East Asian”, “Middle Eastern”, “European”, “African”, “North American”, “South American”, “Hispanic/Latino”, “Indigenous”, “prefer not to answer”, or “other”), their *age* (as an integer) and their *highest level of education* (among “secondary education”, “high school”, “undergraduate degree”, “graduate degree”, “doctorate degree”, “no formal qualifications” and “not applicable”), which might affect the clarity of the texts they write, or how they interpret what they read.

People’s *personality traits* are another attribute that could guide their judgments about mental states. We follow the Big-Five personality measure of Gosling et al. (2003). As an alternative to lengthy rating instruments, it is a 10-item measure corresponding to the dimensions of “openness to experience” (measured positively via “open to new experiences and complex” and negatively via “conventional and uncreative”), “conscientiousness” (measured positively via “dependable and self-disciplined” and negatively via “disorganized and careless”), “extraversion” (measured positively via “extraverted and enthusiastic” and negatively via “reserved and quiet”), “agreeableness” (measured positively via “sympathetic and warm” and negatively via “critical and quarrelsome”), and “emotional stability” (measured positively via “calm and emotionally stable” and negatively via “anxious and easily upset”). Participants self-assign traits by rating each pair of adjectives on a 7-point scale, from “disagree strongly” to “agree strongly”.

As a link between the annotator and the annotation job, we are interested in what *emotion the participants feel* right before entering the questionnaire. For that, the emotion labels presented in Section 3.3.2 need to be scored on a 1–5 scale (i.e., “not at all”, “intensely”), except for the neutral label. Further, we demand that they judge the reliability of their own answers. This variable is instantiated in different ways for the two phases. Since writers can recall events that happened at any point in their life, some memories of appraisals might be more

[4] For the study of neutral events, the emotion duration variable comprises the option “I had none”.

vivid than others, which can affect their annotations. Therefore, we deem *confidence* as the trustworthiness of this episodic memory, quantifying people's belief that what they recall corresponds to what actually happened. In the emotion decoding phase, this variable measures the annotators' confidence that the emotion they recognized is correct. Both are assessed on a 5-point scale, 1 being the lowest degree of confidence.

Lastly, we notice that the goal of building and validating a corpus of self-reports potentially suffers from two major flaws. First, there is no guarantee that the described events happened in the writers' life. Second, the readers' judgments might depend on whether they had an experience comparable to the descriptions they are presented with. Therefore, we ask the writers if they actually experienced the event they recounted, and the validators if they experienced a similar event before. We have no means to assess the honesty of the ratings of this variable. But assuming that the answers can be trusted, *event familiarity* represents an additional level of information for examining patterns of appraisals (e.g., how well the appraisal of events that were not lived in the first person can be reconstructed).[5]

4 Building Event-Centered Corpora

We arranged all variables into various questionnaires, to build crowd-enVENT, deISEAR and enISEAR within our two-phase experimental design. The Phase 1 side of all of them comprises recollections of event experiences; their Phase 2 side is about inferred emotion (and appraisal) evaluations. To a different extent in crowd-enVENT and de(/en)ISEAR, text generation and validation were designed to mirror each other with respect to the considered variables, the formulation of questions, and the possible answers. We now describe the process of creating these corpora as well as the resulting data.

Crowdsourcing details relative to the two studies are in Appendix A, Section 1, which includes: strategies to promote data quality and cost breakdowns for crowd-enVENT (Section 1.3); the full questionnaire to build crowd-enVENT, and comparisons between the generation and the validation phases (Section 1.4); questionnaires, crowdsourcing details and costs for deISEAR and enISEAR (Section 1.5).

[5] Note that *event familiarity* as an additional variable does not correspond to the appraisal *familiarity* (2.) under the Novelty check. The former is an evaluation conducted during the annotation task, the latter in the context of the chosen event.

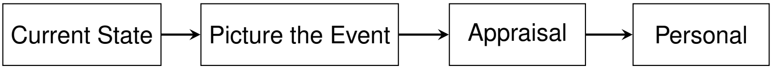


Figure 3.4: Questionnaire overview. The two phases of data creation mainly differ with respect to the block "Picture the Event": in the generation phase, the event is recalled and described; in the successive phase, the validators read the text to put themselves in the shoes of the writers.

4.1 crowd-enVENT

Let us focus on the emotion- and appraisal-motivated annotation activity. This first study can be conceptualized as the bottom portion of Figure 3.1 (Page 77). It is a study on the production and recognition of event *evaluations*. The model in the figure assumes that discrete emotions are inferred by observers (i.e., in our textual setup, the readers) through appraisals (see dashed lines). Our goal is not to test that assumption, but we take on the idea that both appraisals and emotions are variables of interest.

4.1.1 Data Collection

To bridge theoretical insights from psychology and computational emotion analysis, we aimed at creating guidelines (mostly based on the ISEAR questionnaire) that correspond in the two phases as much as possible. We recruited contributors on Prolific[6], targeting speakers of English as a first language.

Phase 1: Event Generation. In the generation phase, each questionnaire was dedicated to a different prompting emotion E (the twelve emotions in Section 3.3.2 and *noemotion*), but all of them instantiated the same template. As shown in Figure 3.4, we collected four blocks of information. In the beginning, participants were asked about their *current emotion state*. They then addressed the task of recalling a real-life event in which they felt emotion E, indicating the *duration of the event*, the *duration of the emotion*, the *intensity* of the experience, and their *confidence*. This generation of descriptions was formulated as a task of sentence completion, observing that this strategy made the job easier for laypersons, without inducing any restriction on sentence structure: they completed the sentence *"I felt E when/because..."* (they saw *"I felt no particular emotion when/because..."* in the *noemotion*-related

[6] https://www.prolific.co.

questionnaire). We encouraged them to write about any event of their choice, and to recount a different event each time they took our survey, in case they participated multiple times. Moving on to the third block of information, people rated the 21 appraisal dimensions, considering the degree to which each of the corresponding statements held for the described event. The survey concluded with a group of questions on demographic information, *personality traits* and *event familiarity*.[7] People who participated repeatedly needed to provide their demographics and personality-related data only once.

We built the data in 9 rounds, and we ensured the same number of descriptions for all emotions, except for *shame* and *guilt*: for each we gathered half the items than for the other emotions, motivated by the affinities between the two (Tracy and Robins, 2006) and the possible difficulty for crowdworkers to discern them.

After the first three rounds, a substantial number of participants had mentioned similar experiences. For instance, *sadness* triggered many descriptions of loss or illness, and *joy* tended to prompt texts about births or successfully passed exams. As we incurred the risk of collecting over-repetitive appraisal combinations, we promoted data diversity from round 4 on. We re-shaped the text production task with two approaches. One served to stimulate the recalling of idiosyncratic facts. The questionnaires based on this solution invited people to talk about an experience that was special to them – one that other participants unlikely had in their life. The other strategy refrained them from talking about specific events. We manually inspected the texts generated up to that point, and compiled a repertoire of recurring topics, emotion by emotion (see Table 3.1); hence, we presented the new participants with the topics usually prompted by E, and we asked them to write something different. This strategy appeared to diversify the data more than the other: we used only this one in the last three rounds, each time updating the list of off-limits topics.

We acknowledge that producing texts by filling in a partial sentence and by having restrictions on certain events is an artificial setup. At the same time, constraining linguistic spontaneity resulted in high-quality data. Compared to a free-text approach, the sentence completion framework represented a way to reduce both the occurrence of ungrammaticalities and the need for writers to mention emotion names – which we would have needed to remove for the validation phase. Moreover, the descriptions so obtained display constructs that are similar to written

[7] *Event familiarity* was included from round 5 afterwards.

Emotion	Off-limits topics
Anger	reckless driving, breaking up, being cheated on, dealing with abuses and racism
Boredom, No Emo.	attending courses/lectures, working, having nothing to do, standing in cues/waiting, shopping, cooking/eating
Disgust	vomit, defecation, rotten food, experiencing/seeing abusive behaviors, cheating
Fear	being home/walking alone (or followed by strangers), being involved in accidents, loosing sights of own kids/animals, being informed about an illness, getting on a plane
Guilt, Shame	stealing, lying, getting drunk, overeating, and cheating
Joy, Pride, Relief	birth events, passing tests, being accepted at school/for a job, receiving a promotion, graduating, being proposed to, winning awards, team winning matches
Sadness	death and illness, loosing a job, not passing an exam, being cheated on
Surprise	surprise parties, passing exams, getting to know someone is pregnant, getting unexpected presents, being proposed to
Trust	being told/telling secrets, opening up about mental health

Table 3.1: Fully-updated list of off-limits topics used to induce event variability.

productions on digital communication channels (e.g., those that can be found in the corpus by Klinger et al. (2018)).

Having concluded all rounds, we compiled the generation side of crowd-enVENT. In total, we obtained 6600 event descriptions balanced by emotion: 275 descriptions for *guilt* and *shame*, and 550 for all other prompting emotion labels.

Phase 2: Emotion and Appraisal Validation. For the second crowdsourcing phase, part of the generated texts were re-annotated by readers. We sampled them with heuristic- and random-based criteria: the data was balanced by emotion (100 per label, except for *guilt* and *shame*, each of which received half the items), and it was extracted from the answers of different generators to boost the linguistic variability shown to the annotators – assuming that personal writing styles emerged from the descriptions. From a set of texts that respected these conditions, we randomly extracted 1200.

Questionnaires here were not dedicated to one predefined prompting emotion. Each of them included 5 texts that could be related to any of the emotions from Phase 1. In the texts, we replaced words corresponding to an emotion name with three dots (e.g., *"I felt ... when I passed the exam"*), to ensure that the validators did not see any explicit emotion, so that the emotion reconstruction task would not be trivial. This preprocessing step was accomplished through rules to mask all words in an E-related text with the same lemma as E or synonyms of E (e.g., the word *"furious"* in texts prompted by *anger*), and heuristics to remove emotion words that contained typos.

The questionnaire template followed that for generation (sketched in Figure 3.4), with a few adjustments and most answering options mirroring those used before. The block of questions opening the survey asked people to rate their *current emotion state*. Next, annotators were presented with one description. They were asked to put themselves in the shoes of the writer at the moment in which she experienced the event, and to attempt and infer the emotion that she was elicited by it. We chose to work in a mono-label setting for compliance with the framework of Scherer and Wallbott (1997). Although their ISEAR corpus only contains the annotations of writers, our validation step instantiates an opposite but corresponding task (i.e., emotion decoding). Thus, we put the readers in the position to provide their predominant impression about E, as were the participants in the previous (emotion encoding) phase. The alternative of picking multiple emotion alternatives for a text might have changed the annotation of the related appraisals, making crowd-enVENT and previous studies on the emotion–appraisal relationship incomparable. The validators also estimated the *duration of the event*, the *duration of the emotion*, as well as the *intensity* of such experience. They rated the *confidence* of their annotations up to that point (i.e., how well they believed they assessed emotion, *event duration, emotion duration*, and *intensity*). As for the variable of *event familiarity*, we asked workers if they had ever had an experience comparable to the one they judged. After that, they reconstructed the original appraisals of the writers.

Participants repeated these steps (included in "Picture the Event" and "Appraisal" in Figure 3.4) consecutively for the 5 texts. Lastly, they provided personal information related to their *age, gender, education, ethnicity* and *personality traits*.

People could take our study only once, such that the judgments of

				event duration					emotion duration					
Emotion	#T	$\overline{\#s}$	$\overline{\#t}$	s	m	h	d	w	s	m	h	d	w	I
Anger	550	1.3	21.8	69	202	107	68	104	16	108	142	114	170	4.2
Boredom	550	1.4	20.4	3	105	306	48	88	6	123	297	53	71	3.6
Disgust	550	1.4	20.6	145	238	58	44	65	30	154	133	97	136	4.1
Fear	550	1.4	22.4	97	233	105	46	69	16	142	143	112	137	4.5
Guilt	275	1.3	21.9	45	92	62	28	48	9	34	55	58	119	4.0
Joy	550	1.3	19.4	61	156	189	65	79	7	57	150	150	186	4.3
No Emo.	550	1.3	17.2	73	256	125	42	54	66	106	65	22	13	2.1
Pride	550	1.3	19.0	67	186	137	49	11	11	54	134	171	180	4.2
Relief	550	1.4	21.7	78	175	140	74	83	32	101	155	121	141	4.3
Sadness	550	1.4	20.7	55	142	111	85	157	7	27	76	112	328	4.5
Shame	275	1.3	20.6	37	114	59	24	41	1	32	65	74	103	4.1
Surprise	550	1.2	18.4	110	235	97	51	57	29	107	153	129	132	4.1
Trust	550	1.3	22.4	35	203	153	61	98	15	93	136	93	213	4.0
$\sum$/Avg.	6600	1.3	20.4	67.3	179.8	126.8	52.7	81.1	18.8	87.5	131.1	100.5	148.4	4.0

Table 3.2: Statistics of the generated data (Phase 1). #T: Number of Texts, $\overline{\#s}/\overline{\#t}$: average number of sentences/tokens. s: seconds, m: minutes, h: hours, d: days, w: weeks. I: Intensity.

each person would appear an equal number of times. Our goal was to obtain a picture of the crowd's impressions appropriate to study inter-subjectivity. Moreover, to prevent writers from annotating their own texts, the study was made inaccessible for all workers who performed generation. Each text was annotated by 5 different people, for a total of 6000 collected judgments (i.e., 1200 texts×5 annotations) with the same amount of judgments per prompting emotion.

4.1.2 Data Analysis

Focusing on the data from Phase 1, we provide descriptive statistics for crowd-enVENT. We then observe patterns of appraisals across emotions, to verify if they align with the insights of studies of self-reports in psychology, which were not conducted via crowdsourcing.

Descriptive Statistics. Table 3.2 illustrates features of the descriptions (for a summary of their semantic content, see Section 1.1 in Appendix B). The corpus contains 6600 texts, 550 per emotion, except for *guilt* and *shame*, having 275 items each. A text consists of one or more sentences.

As shown in column $\overline{\#s}$, the average number of sentences is similar across emotions. Texts are also consistent in terms of length (see $\overline{\#t}$). They comprise 20.43 tokens on average, with *fear* and *trust* receiving the longest descriptions (avg. 22.36) and *surprise* the shortest (avg. 18.38). Non-emotional expressions have fewer words overall, indicating that annotators provided less context to communicate non-affective content. In total, the corpus encompasses 134,851 tokens, excluding punctuation.[8]

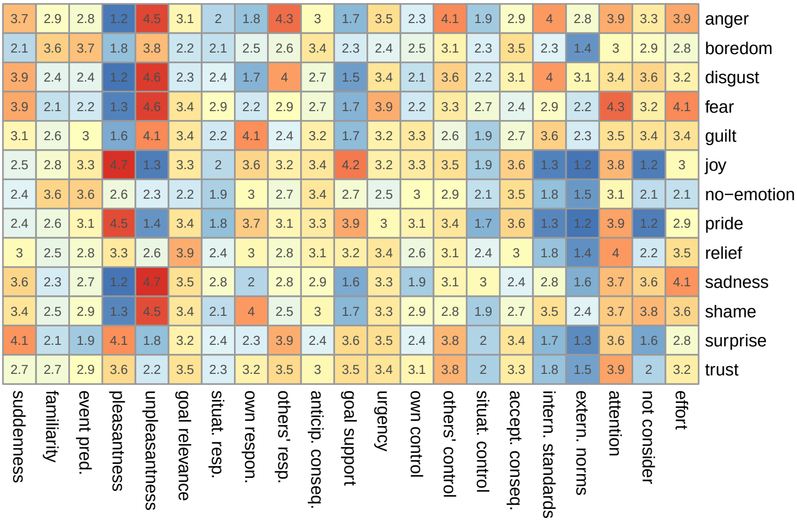
Most texts describe events that took place within minutes or hours ("event duration" in Table 3.2). By contrast, *sadness* has an outstandingly high number of week-long events, and *surprise* and *fear* are characterized by a substantial amount of events that lasted only a few seconds. Interestingly, many texts report on emotions that persisted over days or weeks ("emotion duration"). This collides with the view that emotions are short-lived episodes (Scherer, 2005), but it is unsurprising in our annotation setup. The annotators might have recalled longer emotion episodes in greater detail, and therefore, they might have recounted those to focus on a vivid memory. They might also have perceived long-lasting emotional impacts as being of particular importance (i.e., as special circumstances fitting one of our text diversification strategies).

Another probable criterion by which they picked an episode from their past was the emotion *intensity* connected to it (column "I" in the table): for all labels but *boredom* and *noemotion*, intensity is high. As for the scores of *confidence*, writers generally trusted their memory about the events they described, with average self-assigned confidence above 4.4 across all emotions. The confidence of readers about their own performance is lower, ranging between 3.4 for the *noemotion* instances and 4.1 for *joy*, with an average of 3.9.

Note that besides confidence, crowd-enVENT has other annotation layers that are not reported in the table. They are not central to this study on emotions and appraisal recognition, but some of them (i.e., *emotion state* at present, *event familiarity*, *personality traits*, *age*, *gender*, *ethnicity* and *education*) will be used in Chapter 4, where their distribution in crowd-enVENT will be detailed.

Relation between Appraisals and Emotions. Figure 3.5 shows the distribution of appraisals across emotions as it emerges from the judgments of the writers. Each cell reports the value of an appraisal dimension (on the columns) averaged across all descriptions prompted by a

[8] Tokenization via *nltk*, `https://www.nltk.org`.



	suddenness	familiarity	event pred.	pleasantness	unpleasantness	goal relevance	situat. resp.	own respon.	others' resp.	anticip. conseq.	goal support	urgency	own control	others' control	situat. control	accept. conseq.	intern. standards	extern. norms	attention	not consider	effort
anger	3.7	2.9	2.8	1.2	4.5	3.1	2	1.8	4.3	3	1.7	3.5	2.3	4.1	1.9	2.9	4	2.8	3.9	3.3	3.9
boredom	2.1	3.6	3.7	1.8	3.8	2.2	2.1	2.5	2.6	3.4	2.3	2.4	2.5	3.1	2.3	3.5	2.3	1.4	3	2.9	2.8
disgust	3.9	2.4	2.4	1.2	4.6	2.3	2.4	1.7	4	2.7	1.5	3.4	2.1	3.6	2.2	3.1	4	3.1	3.4	3.6	3.2
fear	3.9	2.1	2.2	1.3	4.6	3.4	2.9	2.2	2.9	2.7	1.7	3.9	2.2	3.3	2.7	2.4	2.9	2.2	4.3	3.2	4.1
guilt	3.1	2.6	3	1.6	4.1	3.4	2.2	4.1	2.4	3.2	1.7	3.2	3.3	2.6	1.9	2.7	3.6	2.3	3.5	3.4	3.4
joy	2.5	2.8	3.3	4.7	1.3	3.3	2	3.6	3.2	3.4	4.2	3.2	3.3	3.5	1.9	3.6	1.3	1.2	3.8	1.2	3
no−emotion	2.4	3.6	3.6	2.6	2.3	2.2	1.9	3	2.7	3.4	2.7	2.5	3	2.9	2.1	3.5	1.8	1.5	3.1	2.1	2.1
pride	2.4	2.6	3.1	4.5	1.4	3.4	1.8	3.7	3.1	3.3	3.9	3	3.1	3.4	1.7	3.6	1.3	1.2	3.9	1.2	2.9
relief	3	2.5	2.8	3.3	2.6	3.9	2.4	3	2.8	3.1	3.2	3.4	2.6	3.1	2.4	3	1.8	1.4	4	2.2	3.5
sadness	3.6	2.3	2.7	1.2	4.7	3.5	2.8	2	2.8	2.9	1.6	3.3	1.9	3.1	3	2.4	2.8	1.6	3.7	3.6	4.1
shame	3.4	2.5	2.9	1.3	4.5	3.4	2.1	4	2.5	3	1.7	3.3	2.9	2.8	1.9	2.7	3.5	2.4	3.7	3.8	3.6
surprise	4.1	2.1	1.9	4.1	1.8	3.2	2.4	2.3	3.9	2.4	3.6	3.5	2.4	3.8	2	3.4	1.7	1.3	3.6	1.6	2.8
trust	2.7	2.7	2.9	3.6	2.2	3.5	2.3	3.2	3.5	3	3.5	3.4	3.1	3.8	2	3.3	1.8	1.5	3.9	2	3.2

Figure 3.5: Average appraisal values as found among the writers' judgments for each emotion. Numbers range between 1 (dark blue) and 5 (dark red).

given emotion (on the rows). High numbers indicate that the appraisal and emotion in question are strongly related. Low values tell us that the appraisal hardly holds for that affective experience.

These results are not only intuitively reasonable but also in line with past studies in psychology (cf. Smith and Ellsworth, 1985). Extreme *suddenness* is related to *surprise*, *disgust*, *fear*, and *anger* more than to other emotions. Instead, *familiarity* commonly holds for events associated with *noemotion* and *boredom*. Another dimension that stands out for these two labels is *event predictability*: its values are comparable to *familiarity* across all emotions, except for *surprise* and *anger*, where it is lower. As expected, *pleasantness* and *unpleasantness* are high for positive emotions (i.e., *joy*, *pride*, *trust*) and negative ones (e.g., *sadness*, *shame*), respectively. Among the positive categories, *trust* has the highest *unpleasantness* value. Also *internal standards* and *external norms* discriminate positive from negative classes, with some within-emotion differences (events sparking negative emotions, e.g., *disgust*, are deemed to violate self-principles more than social norms).

Next, *boredom* and *disgust* are associated with low values for the *goal relevance* of events, while the combination of the three responsibilityoriented appraisals distinguishes a set of emotions: *anger*, *disgust*, and *surprise* stem from events initiated by others (*other responsibility* > *situational responsibility* > *own responsibility*), *guilt* and *shame* are attributed to the self (*own responsibility* > *other responsibility* > *situational responsibility*) and so are *joy* and *pride*, although to a lower degree. Once more, *trust* differs from the other positive emotions, as it accompanies events triggered by other individuals or by the experiencers themselves (e.g., lending someone a precious object) but not by chance. It is interesting to compare the responsibility-specific annotations of *guilt* and *shame* to the three dimensions focused on one's ability to influence events. Also there, the writers felt that the development of the facts was in their *own control* more than in the hands of external factors (*others' control*/*situational control*). Among the two, however, *own control* is especially related to *guilt*, an emotion stemming from behaviors that can be regulated rather than from stable traits of the experiencer (which contribute instead to episodes of *shame* (Tracy and Robins, 2006)). The *anticipation of consequences* reaches particularly low values for *surprise*, *disgust*, and *fear*, with the latter being characterized by the strongest level of *effort* (together with *sadness*) and of *attention*, as opposed to *shame*, *disgust*, and *sadness*, for which the texts' authors reported their attempt to dismiss the event.

These patterns characterizing crowd-enVENT will be resumed later in the chapter to answer our research questions. Before that, the remainder of the section completes the presentation of our resources, drifting attention on enISEAR and deISEAR.

4.2 enISEAR and deISEAR

The creation of crowd-enVENT renders qualitative evidence that crowdsourcing is viable to collect large amounts of appraisal-based data. Yet, further investigation are needed to attest that all such dimensions are beneficial to study annotations in emotion analysis. Therefore, as I move forward to the multilingual setup, I only focus on emotions. Based on discrete models, this study can be conceptualized as the top portion of Figure 3.1.

4.2.1 Data Collection

With deISEAR, we started putting a remedy to the shortage of resources in German. However, only collecting texts in this language would not have enabled us to compare the quality of our resource to that of ISEAR, our point of reference from psychology: we needed to tease apart the effects of the change of setup (in lab to crowdsourcing) and change of language. For that, we collected data in English as well (creating enISEAR).

In both languages, we shortened and adapted the questionnaire of crowd-enVENT to a simplified setting, and hosted our survey on the Figure-Eight[9] crowdsourcing platform.

Phase 1: Event Generation. We asked participants to describe an event in which they felt one of the seven emotions of Scherer and Wallbott (1997) (i.e., *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness*, *shame*[10]). The generation of texts was again formulated as a task of sentence completion (e.g., *"Ich fühlte Freude, als/weil…"*). Writers further specified the *temporal distance*[11] of the event (i.e., whether the event took place "days", "weeks", "months", or "years" before the time of text production), *intensity*, and *emotion duration*. The last two variables are present in ISEAR and crowd-enVENT as well. Compared to these two, here we reduced the answer options from a 5-point scale to 4 alternatives (i.e., for *intensity*: "not very intense", "moderately intense", "intense" and "very intense"; for *duration*: "a few minutes", "one hour," "multiple hours", or "more than one day"). In addition, we asked people to indicate their *gender* ("male", "female", "other").

To obtain enISEAR, we crowdsourced the same set of questions in English (e.g., for the sentence completion task, *"I felt joy when/because…"*).

Phase 2: Emotion Validation. To verify the extent to which the texts convey the *E* for which they were produced, we presented a new set of annotators with ten randomly sampled descriptions (omitting emotion words, e.g., *"I felt … when I found some money in the street."*), together with the list of seven emotions from which they could pick one. Each description was judged by 5 annotators, this time only regarding emotions.

[9] Now "appen": `https://appen.com`.

[10] English→German translations: *anger–Wut, disgust–Ekel, fear–Angst, guilt–Schuld, joy–Freude, sadness–Traurigkeit shame–Scham.*

[11] This variable is in ISEAR but not in crowd-enVENT.

			Temp. Distance				Intensity				Duration				Gender		
		#tok	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	15.1	46	25	31	41	3	25	67	48	23	29	39	52	112	31	–
	Disg.	13.1	38	38	42	25	12	52	48	31	95	37	8	3	110	33	–
	Fear	14.0	25	32	37	49	4	24	58	57	50	32	31	30	109	34	–
	Guilt	13.8	36	27	30	50	8	57	54	24	41	29	43	30	116	27	–
	Joy	11.6	40	30	29	44	2	18	60	63	14	18	42	69	107	35	1
	Sadn.	11.5	29	26	42	46	3	31	43	66	16	9	27	91	113	30	–
	Shame	13.2	25	28	36	54	24	56	41	22	72	28	24	19	116	27	–
	Sum	13.2	239	206	247	309	56	263	371	311	311	182	214	294	783	217	1
English	Anger	28.3	45	29	25	44	9	34	48	52	30	23	36	54	62	81	–
	Disg.	22.4	57	25	21	40	12	51	37	43	66	27	24	26	57	86	–
	Fear	27.0	19	29	36	59	2	30	57	54	52	29	35	27	66	77	–
	Guilt	25.5	33	24	27	59	25	52	43	23	26	39	28	50	59	84	–
	Joy	23.6	32	24	31	56	2	27	48	66	14	13	43	73	60	83	–
	Sadn.	21.6	40	24	31	48	10	45	38	50	17	21	23	82	62	81	–
	Shame	24.8	21	22	19	81	16	51	42	34	29	25	39	50	57	86	–
	Sum	24.7	247	177	190	387	76	290	313	322	234	177	228	362	423	578	–

Table 3.3: Statistics across prompting emotions: average number of tokens (#tok) and extralinguistic labels of the descriptions. Temporal Distance, Intensity and Duration report the number of descriptions for events which took place days (D), weeks (W), months (M) or years (Y) ago, with a specific emotion intensity (NV: not very intense, M: moderate, I: intense, VI: very intense) and duration (min: a few minutes, one hour: h, multiple hours: >h, one or multiple days ≥d); Gender counts of the generators are in the last column (male: M, female: F, other: O). Sadn.: Sadness. Disg.: Disgust.

4.2.2 Data Analysis

What interests about these two corpora is their comparison. We observe them in parallel quantitatively and qualitatively (more analyses are in Appendix B, Section 1.2 and 1.3).

Descriptive Analysis. Both deISEAR and enISEAR comprise 1001 event descriptions. deISEAR includes 1084 sentences and 2613 distinct tokens; enISEAR contains 1366 sentences and a vocabulary of 3066 terms. Table 3.3 summarizes the Phase 1 annotations for each prompting emotion E, reporting the average description length, the annotators' *gender*, and the *duration, intensity* and *temporal distance* of the emotional events.

The main difference between the two languages is description length: English instances are almost twice as long (24.7 tokens) as German ones (13.2 tokens). There are also differences in *gender* distribution, but most patterns are similar across German and English. In both, *anger* and *sadness* receive the longest and shortest descriptions, respectively. Enraging facts are usually depicted through the specific aspects that irritated their experiencers, like *"when a superior at work decided to make a huge issue out of something very petty just to [...] prove they have power over me"*. In contrast, sad events are reported with fewer details, possibly because they are often conventionally associated with pain and require little elaboration, such as *"my grandmother had passed away"*.

Also the perceptual assessments of emotion episodes, as given by the extralinguistic labels of *temporal distance*, *intensity* and *emotion duration*, are comparable between languages. The majority of descriptions are located at the high end of the scale both for *intensity* and *temporal distance*, i.e., they point to "milestone" events that are both remote and emotionally striking.

Post-hoc Event Type Analysis. To better investigate the texts of enISEAR and deISEAR, we manually annotated a handful of features for a sample of 385 English and 385 German descriptions, balanced across emotions. We observed whether a text suggested that the event was reoccurring (*general*); whether the event was actually *past* or expected in the *future*; whether the description alluded to a *prospective* emotion (the experiencer was about to feel) or one already felt; whether the event had a *social* characteristic (involving other people or animals); whether it had *consequences for the self* or *consequences for others*.

These dimensions fit the OCC model of Ortony et al. (1988), where important traits of emotions are their temporality (e.g., *hope* is the prospect of a desirable event), the individuals who performed certain actions, and those touched by their consequences (cf. Figure 2.2 in Chapter 2). These are structural aspects of situations (whereas in the component process theories, evaluations are conditions spurring emotions). Therefore, they appeared appropriate to qualify general semantic features of many of our events. We extended them with other dimensions kin to the consequence-related ones, namely *situational control* and *own responsibility*/*own control* (i.e., appraisal dimensions included in crowd-enVENT), by also annotating whether external factors presumably held *control* over the event or the author held *control* or *responsibility* in the described situation. The guidelines we followed are

	German							English						
Dimension	Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame	Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame
General event	4	2	1	0	0	1	0	2	2	2	2	0	3	0
Future event	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Past event	51	53	53	55	55	54	55	53	53	53	53	55	52	55
Prospective	1	0	4	0	1	1	0	0	0	14	0	1	0	0
Social	30	28	24	29	24	40	25	50	37	30	41	39	49	41
Self conseq.	37	34	37	26	44	21	37	29	26	42	20	35	16	32
Conseq. oth.	21	9	19	34	16	34	14	29	23	19	34	24	43	29
Situat. control	2	5	4	24	9	3	19	3	7	8	31	15	2	24
Own resp./Own control	20	31	17	51	26	23	40	13	29	34	53	34	16	43

Table 3.4: Event type analysis. Cells are counts of post-annotations of 55 descriptions per emotion. Self conseq./Conseq. oth.: consequences for the self/others. Own resp.: own responsibility.

in Appendix A, Section 1.6.

Table 3.4 shows the results. In both English and German, only a few units depict *general* and *future* events, in line with the annotation guidelines. Most descriptions involve other participants, especially in English. Events of *joy* and *fear* seem to have consequences for the authors themselves more than other people. Instead, the majority of situations provoking *guilt* and *sadness* have effects that mainly bear down on others. Regarding *situational control*, *shame* and *guilt* dominate. Like in crowd-enVent, *guilt* seems more frequent in events in which the author is (presumably) *responsible*.

The Emotions of "Others". Our annotations account for *one* emotion (i.e., the writers'), but the post-hoc analysis showed that de(/en)ISEAR spans multiple entities, as many have a *social* trait. Those texts evoke specific situational perspectives for various emoters: an event can involve, affect, or be caused by others beyond the writer. Take the description "*[...] my sister, who is four years younger than me and very spoiled, shouted at my mother for no apparent reason. The heart-broken expression on my mother's face made me feel very angry*". The writer was angered when she perceived the reaction of the mother, who was "heartbroken" in consequence of the shouting; and the sister, in turn, might have had yet another response to her own action (*noemotion*?, *guilt*?). In other words, each participant in an event can appraise it differently.

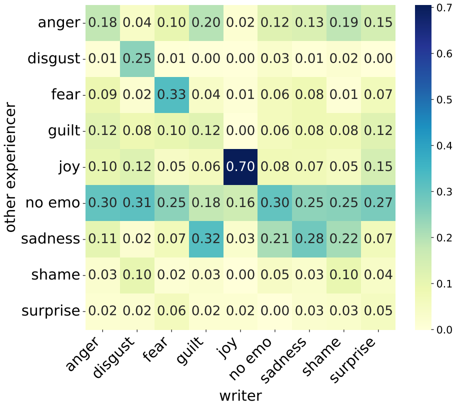


Figure 3.6: Averaged emotion co-occurrences between writers (columns) and other experiencers (rows).

As an experiencer feels a given emotion in an event, what emotions are elicited in the others? We analyzed the link between the emotion and the appraisals of the writer vs. any other named entity in a small-scale study. Four in-house coders annotated emotions and appraisal dimensions in enISEAR performing a task comparable to semantic role labeling. They identified all entities involved in an event, and indicated if and what emotion the entities felt. For each entity, they selected the portion of text that referred to the salient trait of the event sparking that emotion in her, and scored her likely appraisals.

The emotion-centered results are in Figure 3.6, where one cell represents the proportion of times any experiencer is attributed the emotion on the rows, when the writer is annotated with the emotion on the columns.[12] While the same emotion is often common to different experiencers (see the comparably high values on the diagonal), the high numbers scattered off-diagonal indicate that different emotion reactions can be inferred from text for different semantic roles. The writers' *guilt* is often accompanied by another's *sadness* (.32); their *shame* often co-occurs with the *anger* of third parties (.19); interesting combinations are also *guilt–anger* (.20) *noemotion–sadness* (.21). The mentioned entities are often not attributed any emotion in situations that, instead, caused some in the writer (cf. row *noemotion*). In sum, this finer-grained emotion analysis brings to light richer affective reactions than those of a validation setting that only decodes the writers' states.

Also the entity-specific appraisal analysis reveals interesting

[12] These numbers differ from those in the corresponding publication (Troiano et al., 2022a) because the latter includes annotations of units extracted also from ISEAR, Empathetic-Dialogues (Rashkin et al., 2019) and Event2Mind (Rashkin et al., 2018). Here I only consider the post-annotations of enISEAR.

between-experiencer patterns in enISEAR. As one experiencer perceives, for instance, *self responsibility*, what is the appraisal of another participant in the same event? Examples are: *internal standards* of the writer is positively correlated with a strong attribution of *external norms* for others (Spearman's ρ = .67) and vice versa (ρ = .50); *others' control* attributed to the writer tends to accompany *own control* for another experiencer (ρ = .46). A detailed between-experiencer and within-experiencer analysis is in Appendix B, Section 1.6, which also compares the writers' appraisals (as reconstructed by readers) between crowd-enVENT and enISEAR.

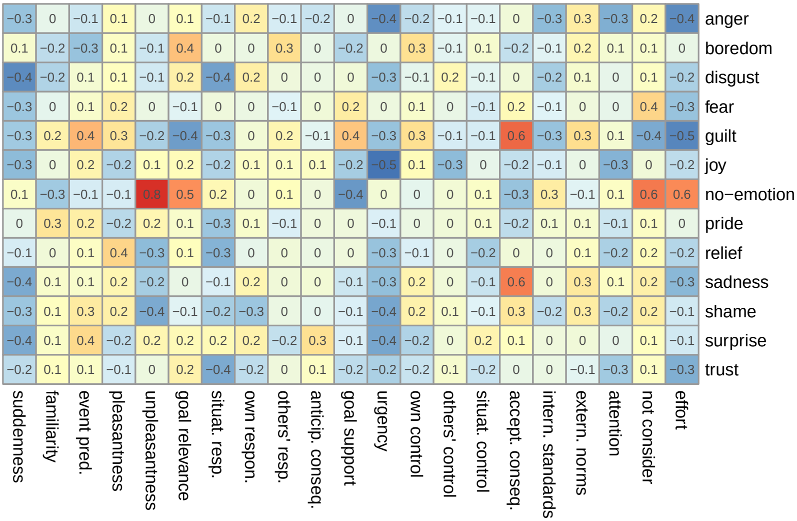
5 Emotion and Appraisal Recognition

We now observe how the readers performed their validation task in crowd-enVENT, deISEAR and enISEAR, to understand how well emotions (and appraisals) can be inferred from factual statements that encode emotions implicitly. This constitutes the core investigation of the chapter that answers our research questions. We incorporate it with an additional analysis concerning automatic emotion detection.

5.1 Human-based Classification

The availability of ground truth and validation labels can be used to answer RQ1 (How well do people infer emotions from implicit expressions?) and RQ2 (How do they infer appraisals from text?) with an investigation of the judgments collected in Phase 2 against those in Phase 1 for either variable.

We answer RQ1 and RQ2 with crowd-enVENT in Section 5.1.1. First, we sharpen the focus on appraisals and their link to emotions (comparing it as found among generators and validators) and we compute inter-annotator agreement. Next, to better grasp the value of event evaluations for data collection studies, we inspect instances in which the validators were either particularly successful or unsuccessful in recovering the writers' emotions and/or appraisals. With this qualitative analysis that identifies some patterns of judgment, we answer RQ3 (Do appraisal judgments help understand variations among emotion assignments?). Section 5.1.2 resumes RQ1 and answers it with the emotion judgments of enISEAR and deISEAR. To conclude the human-based analysis, Section 5.1.3 gives notice of how to annotate appraisals



	suddenness	familiarity	event pred.	pleasantness	unpleasantness	goal relevance	situat. resp.	own respon.	others' resp.	anticip. conseq.	goal support	urgency	own control	others' control	situat. control	accept. conseq.	intern. standards	extern. norms	attention	not consider	effort
anger	−0.3	0	−0.1	0.1	0	0.1	−0.1	0.2	−0.1	−0.2	0	−0.4	−0.2	−0.1	−0.1	0	−0.3	0.3	−0.3	0.2	−0.4
boredom	0.1	−0.2	−0.3	0.1	−0.1	0.4	0	0	0.3	0	−0.2	0	0.3	−0.1	0.1	−0.2	−0.1	0.2	0.1	0.1	0
disgust	−0.4	−0.2	0.1	0.1	−0.1	0.2	−0.4	0.2	0	0	0	−0.3	−0.1	0.2	−0.1	0	−0.2	0.1	0	0.1	−0.2
fear	−0.3	0	0.1	0.2	0	−0.1	0	0	−0.1	0	0.2	0	0.1	0	−0.1	0.2	−0.1	0	0	0.4	−0.3
guilt	−0.3	0.2	0.4	0.3	−0.2	−0.4	−0.3	0	0.2	−0.1	0.4	−0.3	0.3	−0.1	−0.1	0.6	−0.3	0.3	0.1	−0.4	−0.5
joy	−0.3	0	0.2	−0.2	0.1	0.2	−0.2	0.1	0.1	0.1	−0.2	−0.5	0.1	−0.3	0	−0.2	−0.1	0	−0.3	0	−0.2
no−emotion	0.1	−0.3	−0.1	−0.1	0.8	0.5	0.2	0	0.1	0	−0.4	0	0	0	0.1	−0.3	0.3	−0.1	0.1	0.6	0.6
pride	0	0.3	0.2	−0.2	0.2	0.1	−0.3	0.1	−0.1	0	0	−0.1	0	0	0.1	−0.2	0.1	0.1	−0.1	0.1	0
relief	−0.1	0	0.1	0.4	−0.3	0.1	−0.3	0	0	0	0	−0.3	−0.1	0	−0.2	0	0	0.1	−0.2	0.2	−0.2
sadness	−0.4	0.1	0.1	0.2	−0.2	0	−0.1	0.2	0	0	−0.1	−0.3	0.2	0	−0.1	0.6	0	0.3	0.1	0.2	−0.3
shame	−0.3	0.1	0.3	0.2	−0.4	−0.1	−0.2	−0.3	0	0	−0.1	−0.4	0.2	0.1	−0.1	0.3	−0.2	0.3	−0.2	0.2	−0.1
surprise	−0.4	0.1	0.4	−0.2	0.2	0.2	0.2	0.2	−0.2	0.3	−0.1	−0.4	−0.2	0	0.2	0.1	0	0	0	0.1	−0.1
trust	−0.2	0.1	0.1	−0.1	0	0.2	−0.4	−0.2	0	0.1	−0.2	−0.2	−0.2	0.1	−0.2	0	0	−0.1	−0.3	0.1	−0.3

Figure 3.7: Comparison between the average appraisal values assigned by the generators and the validators, divided by emotion. Cells in the red spectrum indicate that the generators on average picked higher scores, and vice versa for cells with negative numbers (in blue). Zero values indicate a perfect match between the average scores of the two phases.

in the future, in the absence of the writers' ground truth.

5.1.1 Recognition on crowd-enVent

The patterns of appraisals across emotions found among the generators (Figure 3.5) provided a picture of the cognitive dimensions underlying emotions. We inspect the same information here, but including the validation side of crowd-enVent. We compare the two batches of judgments in the heatmap of Figure 3.7. In contrast to Figure 3.5, numbers are here computed considering the 1200 texts that underwent Phase 2, and a cell shows the average difference between the gold standards given by the experiencers and the readers' assessments. Should the validators' appraisals be similar to those of the people who lived through the events (thus approaching 0 throughout Figure 3.7), we could conclude that it is possible to obtain corpora with reliable appraisal labels via traditional annotation methods, based on external readers.

Divergent ratings stand out for *unpleasantness*, *goal relevance*, *not consider* and *effort* in the row *noemotion*, as well as for *urgency* in *joy*, *effort* in *guilt*, and the *accept. conseq.* in both *guilt* and *sadness*. *Suddenness*, *effort* and *urgency* have lower values across all emotions, while for *event predictability*, *external norms*, and *not consider*, the validators tended to choose ratings that surpassed the original ones. Overall, these differences are comparably low (all absolute values are below 1), which suggests that on a large scale, the patterns one can study from the two batches of judgments do not diverge from one another radically.

Inter-Annotator Agreement. To address RQ1 and RQ2 more directly, we discuss inter-annotator agreement among the validators, as is usually done in emotion analysis, and between them and the generators. We take all study participants who generated/validated the same texts and pair them. In total we obtain 6,600 generator–validator (G–V) pairs (each generator is coupled with 5 validators) and 12,000 validator–validator (V–V) pairs ($\binom{5}{2} \cdot 1200$), and compute agreement with those: for emotions, we use average F_1 and accuracy, for appraisal annotations, we employ average RMSE scores. We do not normalize for expected agreement, as is commonly done with κ measures, because we do not have unique annotators that remain stable over a considerable amount of texts – which prevents us from assigning a meaningful value for the expected agreement to each individual. We calculate the statistical significance of our results under a .95 confidence level via bootstrap resampling (1000 samples) on the textual instances, pairwise for all results for each evaluation measure (Canty and Ripley, 2021; Davison and Hinkley, 1997).

We obtain F_1 = 0.49 on emotion recognition with G–V, and F_1 = 0.49 with V–V pairs. Agreements stratified by emotion are reported in Table 3.5. Generators and validators achieve a lower agreement on *anger*, *joy*, *pride*, *trust*. Notably, emotions where G–V agree more (*boredom*, *fear*) are those on which V–V pairs also achieve the highest consistency. The calculation of accuracy returns a slightly different impression, as we find that the difference between G–V (0.50) and V–V pairs (0.52) is significant. The difference in agreement for the annotation of appraisals is also significant, and more noticeable (1.57 for G–V vs. 1.48 for V–V). Table 3.6 reports results by appraisal dimension. The biggest difference holds for *not consider*, followed by *other responsibility*, *situational responsibility* and *suddenness*. In no appraisal dimension do G–V pairs outperform V–V pairs.

	F_1	
	G–V	V–V
Anger	0.47	0.51
Boredom	**0.66**	**0.65**
Disgust	0.57	0.56
Fear	**0.64**	**0.60**
Guilt	0.46	0.44
Joy	0.45	0.53
No-emotion	0.28	0.23
Pride	0.51	0.57
Relief	0.54	0.54
Sadness	0.55	0.59
Shame	0.38	0.33
Surprise	0.35	0.32
Trust	0.47	0.48

Table 3.5: Inter-annotator agreement on emotions in crowd-enVENT. We measure it within phase (V–V) and between phases (G–V) via F_1 score. Emotions are the prompting labels treated as gold standards.

	RMSE		
	G–V	V–V	Δ
suddenness	1.57	1.46	**0.11**
familiarity	1.58	1.47	0.10
event pred.	1.56	1.49	0.06
pleasantness	**1.22**	**1.19**	0.03
unpleasantness	1.33	1.24	0.08
goal relevance	1.62	1.52	0.09
situat. resp.	1.65	1.53	**0.11**
own resp.	1.45	1.40	0.05
others' resp.	1.59	1.45	**0.13**
anticip. conseq.	1.69	1.60	0.08
goal support	1.50	1.40	0.09
urgency	1.76	1.67	0.08
own control	1.56	1.48	0.07
others' control	1.63	1.55	0.07
situat. control	1.65	1.55	0.09
accept. conseq.	1.83	1.73	0.10
int. standards	1.56	1.49	0.07
ext. norms	**1.30**	**1.26**	0.04
attention	1.52	1.52	0.00
not consider	1.64	1.48	**0.16**
effort	1.60	1.51	0.09

Table 3.6: Appraisal inter-annotator agreement measured via average Root Mean Square Error. Δ: difference between G–V and V–V. For consistency, appraisal dimensions are sorted in the same order as in Figure 3.5 and 3.7.

These agreement scores are unimpressive, as one can expect from an emotion analysis task conducted with implicit texts; but focusing on the comparison between the G–V and V–V sides, we derive an insight that answers RQ1: validators agree with the point of view that they are attempting to reconstruct in a similar way as with all other judges undertaking the same task. Despite the significant difference in accuracy, readers do not agree among each other substantially more than with the generators, which suggests that the readers' consistency found in an annotation endeavor can be taken as a good measure also for their correctness.

Results about appraisals are more similar to the insight from psychology according to which inter-annotator agreement changes as one changes the annotators where it is computed (G–V vs. V–V). To answer RQ2, our short texts allow the readers to reconstruct event evaluations, but only to a certain extent. That is unsurprising considering our annotation setup. A reader might correctly estimate an emotion, while not perfectly inferring the original event evaluations, either because she has in general a different approach than the writer to rating items on a 5-point scale, or because the text does not provide sufficient information to reconstruct multiple and fine-grained dimensions, which might lead her to report on a prototypical way of evaluating the event. This motivates us to take a look at inter-annotator agreement qualitatively, and better understand the relationship between emotion and appraisal judgments.

Qualitative Discussion of Agreement. We investigate the texts on which judges (dis)agree and divide them into two curricula, i.e., those that turned out "easy" to label and those where the correct inference was difficult to draw. Table 3.7 shows examples in which all readers correctly reconstructed the writers' emotions. Table 3.8 reports items where all validators inferred the same emotion, but that emotion does not correspond to the gold label – as revealed in the quantitative discussion, agreeing on the emotion does not imply agreeing on the appraisals. We follow this idea by dividing the tables in two blocks. The top block corresponds to texts with high G–V agreement in appraisal (as an average RMSE), while the bottom to high disagreement.

The top examples in Table 3.7 describe events that have unambiguous implications for the well-being of the experiencer. Ranging from ordinary circumstances (e.g., baking) to peculiar ones (e.g., being threatened by a housemate), these texts can be argued to depict facts with shared underlying characteristics, graspable even from people who did not live through them (e.g., most likely, being threatened spurs *unpleasantness*, scarce *goal relevance*, and inability to *anticip. conseq.*). By contrast, the examples with low agreement on appraisals seem to require a more elaborate empathetic interpretation. They might be easily understandable regarding the emotion, but they underspecify many details about the described circumstance, which would be necessary for a reader to infer how it was evaluated along fine-grained dimensions. For instance, going to the hospital is attributed to *fear* (example Id 209), but it remains unclear under which circumstances this situation takes place (a planned surgery? an accident? to visit someone?).

	Emo.		Appr.	
Id	G	V	RMSE	Text
1	pride		0.65	I baked a delicious strawberry cobbler.
2	fear		0.69	I was running away from a shooting and a car was trying to run me down
3	fear		0.72	I felt ... when there was a power outage in my home. That day, my wife and I were cuddling in the sitting room when a thunderstorm started. Then ... filled me when thunder hit our roof and all the lights went off.
4	pride		0.82	I felt ... when I ran a marathon at a decent pace and finished the race in a good place
5	fear		0.84	A housemate came at me with a knife.
6	fear		0.86	I was surrounded by four men; they hit me in the face before I offered to give them everything I had in my pockets.
7	pride		0.89	I felt ... when I accomplish my goals through a team effort. I take part in team sports and have a pivotal role in success, and being able to do my job and make my team proud of me gives me a strong sense of ....
203	fear		1.68	I felt ... when I was in a public place during the coronavirus pandemic
204	pride		1.73	I helped out a friend in need
205	fear		1.74	I felt ... when i had a night terror.
206	boredom		1.81	I went on holiday abroad for the first time. I felt ... because I didn't enjoy being on the beach doing nothing.
207	sadness		1.86	I felt ... when I graduated high school because I remember that I'm growing up and that means leaving people behind.
208	disgust		2.03	His toenails where massive
209	fear		2.08	I felt ... going in to hospital
210	trust		2.35	my husband is always there for me and i can ... that no matter what he will be there for our child and do what ittakes to provide for us as a family

Table 3.7: Examples where all validators (V) correctly reconstructed the emotion of the generators (G). The top (bottom) examples have high (low) agreement on appraisals.

	Emo.		Appr.	
Id	G	V	RMSE	Text
1	joy	pride	0.81	finally mastered a song i was practising on guitar
2	pride	joy	0.83	my band got signed to a label run by an artist i admire
3	trust	joy	0.87	I am with my friends
4	joy	pride	0.90	I bought my own horse with my own money I had worked hard to afford
5	surprise	pride	0.93	when I built my first computer
6	surprise	joy	1.00	I felt ... when my partner put their arms around me at a concert and started to dance with me to a song we listen to.
7	trust	joy	1.01	I felt ... when my boyfriend drove out of town to see me at 2 in the morning.
8	anger	fear	1.09	My waters broke early during pregnancy
9	joy	pride	1.11	I was able to complete a challenge that I didn't think I would do
43	pride	sadness	1.65	That I put together a funeral service for my Aunt
44	surprise	joy	1.66	I got a dog for my birthday
45	joy	relief	1.68	I was diagnosed with PMDD because it meant I had answers
46	no-emotion	anger	1.69	I saw an ex-friend who stabbed me in the back with someone I considered a friend
47	shame	relief	1.81	I tasked with sorting out some files from the office the previous day and I slept off when I got home
48	disgust	sadness	1.82	I was left out of a family chat.
49	sadness	relief	1.83	when I returned to my apartment after being away during COVID.
50	shame	sadness	1.84	Not being around my son
51	surprise	joy	1.90	I found the perfect man for me, and the more time goes on, the more I realized he was the best person for me. Every day is a ....
52	no-emotion	sadness	1.93	Breaking up with my partner

Table 3.8: Examples in which all validators (V) agreed with each other, but not with the generators (G) of the event descriptions. The top (bottom) blocks shows texts where the agreement is high (low) on appraisals.

Table 3.8 contains texts from which readers did not recover the actual emotion experienced by the author. Instances of high appraisal agreement are associated to labels with similar affective meanings, and therefore are more likely to be confused than, e.g., a positive and a negative emotion. The mislabeling mostly happens between *joy* and *pride*, both of which are (arguably) appropriate, and in one case between *anger* and *fear*. Instead, the bottom block of the table reports texts in which a positive emotion is misunderstood for a negative one. For instance, Id 43 was produced for *pride* but was validated as *sadness*. These mistakes might be due to the readers focusing on a portion of text different from that considered salient by the writer (e.g., Id 49, "being away with covid": *sadness*, "returning home": *relief*), or to the readers drawing a presupposition from the text (e.g., Id 43, a funeral took place: *sadness*) different from what the author intended to convey (i.e. he/she was able to organize it: *pride*). It is also possible that some of these G–V disagreements derive from the sequence of tasks in the survey. The readers were first prompted to assign an emotion to the event and only later were they guided to evaluate it in detail. Going the other way around might have led the crowdworkers to reflect on the events in a more structured way, and might have elicited different judgements.

There are also examples in which an emotion is assigned, while none was felt by the event experiencer (e.g., Id 46 and 52). This can be interpreted in various manners. It might signal the subjectivity of emotions, but it also tells something about how some writers tackled the task. When prompted by another E, they might have had expectations about *our* expectation for their productions (e.g., that their text expressed E in an as unambiguous way as possible): they likely resorted to instantiations of E that (they might not have experienced but that) we would have considered appropriate and worth of rewarding. E = *noemotion* raised perhaps less pressing expectations: participants recounted circumstances that usually do not leave individuals in apathy, but that did not perturb their general sense of feeling in that occasion. We thus conjecture that these are the instances on which generators "lied" the less.

Brought together, these observations illustrate features of crowd-enVent, and suggest some systematic patterns in its annotation that are informative about agreement (for RQ3). To begin with, part of the instances that we collected convey enough information for readers to understand emotions, independent of if and how they also understand the underlying evaluation. From this, we derive that *at least in some cases, grasping appraisals from text is not necessary to grasp the corresponding emotion*.

Second, there are instances where humans fail to reconstruct emotions, and differences between such judgments are mirrored in differences in their appraisal measures.

Third, by contrasting the high vs. low appraisal agreements blocks of Table 3.8, we learn that the "semantic difference" between the incorrectly reconstructed emotions is lower if the appraisals are inferred acceptably well (e.g., readers picked *pride* instead of *joy*, while they face confusions between more incongruent labels, e.g., *pride*/*sadness*, by disagreeing also on the appraisals): the annotators can share the underlying understanding of an affective experience, even if they disagree on a discrete label to name it. Hence, the labels they chose can be considered compatible alternatives. As our single-label experimental setup did not request the generators to indicate multiple emotion labels for their experience, a follow-up study would be needed to confirm this hypothesis.

5.1.2 Recognition on deISEAR and enISEAR

Returning to the problem of emotion inferences (RQ1), we analyze to what extent the emotions labeled in Phase 2 agree with the prompting emotion presented in Phase 1 in enISEAR and deISEAR. Earlier, our use of F_1 and accuracy as inter-annotator agreement measures followed other research in emotion analysis (Haider et al., 2020; Štajner, 2021). Their comparison however turned out hard to interpret (only with accuracy were G–V and V–V significantly different). We thus opt for a simpler and more transparent measure here. Table 3.9 reports for how many descriptions (out of 143) the prompting emotion was selected at least one, two, three, four, or exactly five (out of five) times in Phase 2. Agreement is similar between deISEAR and enISEAR. This indicates that the German items, although short, convey as much emotion information as the more elaborate English descriptions. In both languages, the agreement drops across the columns, yet half of the descriptions show perfect inter-subjective validity (=5): 505 for German, 499 for English.

Again, we find differences among emotions. Agreement is nearly perfect for *joy* and rather low for *shame*. These patterns can arise due to different processes. Certain emotions are easier to recognize from

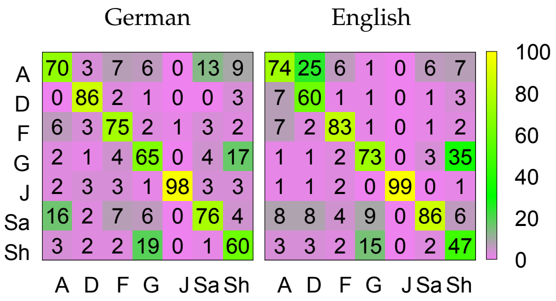
	German					English				
	≥1	≥2	≥3	≥4	=5	≥1	≥2	≥3	≥4	=5
Anger	135	125	107	81	52	137	129	112	89	59
Disgust	139	134	130	124	91	118	101	84	76	53
Fear	134	124	108	99	78	136	131	124	116	86
Guilt	137	126	102	67	31	137	130	124	89	44
Joy	142	142	142	140	136	143	143	143	143	137
Sadness	132	123	113	97	76	140	133	131	116	97
Shame	128	109	86	66	41	116	92	64	41	23
Sum	947	883	788	674	505	927	859	782	670	499

Table 3.9: Human emotion recognition. Numbers are counts of descriptions whose prompting label (on the rows) agrees with the emotion labeled by all Phase 2 annotators (=5), by at least four (≥4), at least three (≥3), at least two (≥2), at least one (≥1).

language (e.g., *"when I saw someone else got stabbed near me"*: *fear*) than others (e.g. *"when my daughter was rude to my wife"*: elicited for *shame*, arguably fitting also *anger* or *sadness*). Patterns may also indicate closer conceptual similarity among specific emotions (Russell and Mehrabian, 1977, i.a.).

To follow up on this observation, Figure 3.8 shows two confusion matrices for German and English. They plot the frequency with which annotators selected emotion labels (Phase 2, rows) for the texts of each prompting emotion (Phase 1, columns). Numbers on the diagonals correspond to the "=5" columns in Table 3.9, mirroring the overall inter-subjective validity of the descriptions, from the highest (on *joy*) to the lowest (on *shame*). The off-diagonal cells indicate disagreements. In both languages, annotators perceive *shame* descriptions as expressing *guilt* and vice versa (35% and 15% for English, 17% and 19% for German). In fact, both *shame* and *guilt* "occur when events are attributed to internal causes" (Tracy and Robins, 2006), and thus they may appear overlapping.

We also see an interesting cross-lingual divergence. In deISEAR, *sadness* is comparably often confused with *anger* (13% of items), while in enISEAR it is *disgust* that is regularly interpreted as *anger* (25% of items). This might results from differences in the connotations of the prompting emotion words in the two languages. For *disgust* (*"Ekel"*), German descriptions concentrate on physical repulsion, while



German

	A	D	F	G	J	Sa	Sh
A	70	3	7	6	0	13	9
D	0	86	2	1	0	0	3
F	6	3	75	2	1	3	2
G	2	1	4	65	0	4	17
J	2	3	3	1	98	3	3
Sa	16	2	7	6	0	76	4
Sh	3	2	2	19	0	1	60

English

	A	D	F	G	J	Sa	Sh
A	74	25	6	1	0	6	7
D	7	60	1	1	0	1	3
F	7	2	83	1	0	1	2
G	1	1	2	73	0	3	35
J	1	1	2	0	99	0	1
Sa	8	8	4	9	0	86	6
Sh	3	3	2	15	0	2	47

Figure 3.8: Emotion confusion matrices. Columns: emotions prompting generation; rows: validated emotions.

the English descriptions also include metaphorical *disgust* which is more easily confounded with other emotions such as *anger*. Further agreement analyses are in Appendix B, Section 1.4.

Overall, these results corroborate our G–V findings on crowd-enVENT for RQ1. Slightly more than half the items are correctly classified by all humans reading them. We take this as a sign of the data quality, and of the inter-subjective validity of those items. However, the other half spurs multiple interpretations different from the ground truth, evidence that implicit expressions are prone to being mistakenly decoded.

5.1.3 Appraisals without Ground Truth

A feature evident in several texts of our three resources is their brevity (e.g., "*I felt . . . when my friend betrayed me*"). On that account, the idea that the 21 appraisals, so diverse and fine-grained, are purely inferred from linguistic material in crowd-enVENT is hardly tenable. The original event evaluations might be signalled from the descriptions at times; but other times judges rely more on their own knowledge to make assumptions about the considered situation.

An auxiliary experiment confirmed indeed that text alone does not always carry sufficient information to correctly score event evaluations. With in-lab readers, we labeled enISEAR in two further rounds of annotation distanced by six months: first in a regular setting, where they attempted to reconstruct the writers' appraisals from the texts, and then by having access to both a text and its prompting emotion label. Knowledge of the emotion boosted the annotator's agreement by a substantial amount (+13 points in Cohen's κ, from .55 to .68).

For future studies, this insight devises a plan to add appraisal labels to pre-existing texts (which lack ground truth) while alleviating the readers' low consistency: if some types of information about internal states are available for the texts (e.g., the writers' emotions), their disclosure can help the coders make hypotheses about other cognitive information. Such an annotation approach can be argued to return a distorted view on humans' ability to infer appraisals, because it facilitates their task; but actually, it still has a realistic blueprint. In everyday communication, we ask our interlocutors what they felt (e.g., surprise) in order to update other and more subtle understandings of how they might have judged a situation (e.g., the interlocutor did not expect it to happen). We can also ask, vice versa, how the situation was judged (e.g., "Did you expect that?") to refine our estimate of their emotional state (e.g., surprise) – whether knowledge of appraisals improves humans' emotion annotation is to be confirmed.

5.2 Machine-based Classification

As a last analysis, we address automatic emotion classification. We have two goals. First, to shed light on the quality and comparability of enISEAR and deISEAR. Second, to assess the usability of crowd-enVENT. In either case, the task is particularly demanding. Inferences must be drawn from implicit expressions out of context, and under the same conditions of the Phase 2 crowdsourcing step: our automatic models estimate a single label from the set of prompting emotions, for texts where emotion words are obscured.

enISEAR and deISEAR are inappropriate to train classifiers because they are limited in size. They are also different in language, which requires either the use of a multilingual model or of two equivalent systems in English and German. For both, we would need to access large resources of event descriptions in German, but to our knowledge none is available. A straightforward solution is translating the data to have all texts in one language, and applying one classifier to those. Translation proved a valuable approach for cross-lingual modeling (Barnes et al., 2016). The lexical and structural changes it produces can lose some affective properties of the texts. For instance, studies in sentiment analysis observed that polarized texts can be neutralized (Petrova and Rodionova, 2016), but we expected minimal "noise" given the shortness of the texts in our resources.

We thus trained a classifier on the original ISEAR produced in lab (7665 instances), which is of relatively high quality, and we evaluated it on all instances collected in Phase 1, both enISEAR and a version of deISEAR translated with Google Translate.[13] We used a Maximum Entropy classifier with L2 regularization and boolean unigram features (in liblinear, Fan et al., 2008) as a comparably strong baseline, easier to reproduce than most neural classifiers due to its convex optimization function and fewer hyper-parameters (other details in Appendix B, Section 1.5). For this experiment, using ISEAR as a training set was advantageous over crowd-enVENT, since ISEAR contains more varied sentence structures (the texts were not written via sentence completion) and less similar events to enISEAR and deISEAR, having been produced further back in time and with a specific sample of contributors (demographics-wise).

Table 3.10 (a) reports the results, showing a decent performance of the model on our novel corpora (micro and macro avg. F_1 = .47 for both). The scores and differences between emotion classes resemble those of previous studies (e.g., Bostan and Klinger, 2018) that did not directly focus on the arguably more challenging case of implicit emotion expressions. Modeling performance and inter-annotator (dis)agreement are correlated: Spearman's ρ between F_1 and the diagonal in Figure 3.8 is 0.85 for German, $p = .01$, and 0.75 for English, $p = .05$. Emotions that are difficult to annotate are also difficult to predict.

It is noteworthy that results for German are on par with English despite the translation step. We take this outcome as evidence for the cross-lingual comparability of deISEAR and enISEAR. However, we have no guarantee that emotion-related information was kept unaltered during translation and that, by applying a classifier learned directly in German, results would have been similar (later, this thesis will investigate emotion transformations in translation). Overall, the acceptable scores obtained in English informed us about the comparability between enISEAR and ISEAR, confirming that crowdsourcing did not sacrifice data quality; observing the English results against those on deISEAR revealed that the change of language did not affect such quality either.

Let us now verify how well a classifier learned on our data, with no recourse to external corpora, infers emotions. As a test set for this experiment, we used the 1200 descriptions of crowd-enVENT that en-

[13] `http://translate.google.com`, applied on February 25, 2019.

(a)

	deISEAR	enISEAR
Anger	.29	.27
Disgust	.49	.45
Fear	.48	.57
Guilt	.42	.41
Joy	.68	.67
Sadness	.53	.58
Shame	.39	.32

(b)

	crowd-enVENT	
	model	humans
Anger	.53	.57
Boredom	.84	.73
Disgust	.66	.65
Fear	.65	.73
Guilt	.48	.53
Joy	.45	.49
No-emotion	.55	.33
Pride	.54	.59
Relief	.63	.64
Sadness	.59	.63
Shame	.51	.48
Surprise	.53	.42
Trust	.74	.52

Table 3.10: Automatic emotion recognition performance (F_1). Experiments on deISEAR and enISEAR (a) are conducted with a bag-of-words Maximum Entropy model trained on ISEAR and tested on our multilingual datasets (for deISEAR, we use an English translation). For the experiment on crowd-enVENT (b), we fine-tune and test a RoBERTa-based model on our data, and compare its performance to humans'.

tered Phase 2. We divided the rest of the generation instances randomly into training and validation data (90 % for training, i.e., 4860 texts; 10 % for validation, i.e., 540 texts). This second model was built on top of a pretrained RoBERTa-large model (Zhuang et al., 2021), adding a classification layer after the average-pooled output representations. Validation data was used to perform early stopping.[14]

We confronted this model with the readers' performance on crowd-enVENT. Hence, to obtain the same level of granularity for humans as for the automatic predictions, we aggregated their judgments. For each instance, we used the majority vote, and whenever that led to a tie, we assigned a higher weight to the judgments of annotators who indicated a strong degree of confidence in the surveys.

Column "model" in Table 3.10 (b) shows results averaged across 5 runs. Other than being substantially higher than those obtained

[14] Details as in Troiano et al. (2023b), where the experiments were conducted by my co-author Laura Oberländer.

by the Maximum Entropy model, the scores (macro avg. F_1 = .59) surpass humans by 3pp[15] (column "humans", macro avg. F_1 = .56). These numbers speak in favor of the quality of crowd-enVENT, but they are just preparatory for future research because certain types of events occur multiple times in crowd-enVENT (cf. our text induction strategies), while we did not enforce event variability between the training and testing sets. Hence, more careful investigation is necessary to determine how a classifier behaves with never-before-seen events.

Establishing a fairer comparison between models and validators would also be advisable. Strictly speaking, the two did not undertake the same task: the models were trained on the writers' labels, the readers attempted to minimize the distance between their own point of view, prior experiences and subjective interpretations, and that of (to them) unknown text author. Moreover, following the assumption that humans leverage appraisals to sense the emotions of others, classification tasks could make use of these evaluative variables. Initial findings are in Troiano et al. (2023b), where appraisal dimensions were used as input features together with text, and they proved to boost text-based emotion classification, to an extent. Appraisals also gave a tool to introspect the decisions of the models (e.g., some dimensions played a role in resolving misclassifications based on text alone). However, computational emotion analysis still lacks robust evidence to conclude that appraisals improve this automatic task.

6 Discussion and Conclusion

Aimed at studying humans' emotion recognition ability from implicit expressions, this chapter presented a twofold contribution. From a practical standpoint, I curated resources (now publicly available[16]) of event descriptions. Second, I proposed a novel paradigm for emotion analysis comprising 21 dimensions motivated by theories from psychology.

crowd-enVENT, deISEAR, and enISEAR rest upon a crowdsourcing schema that reflects the flow of emotion production–decoding in a simple communication framework. They are all labeled by external

[15] Being based on judgments aggregation, humans' F_1 scores are different from those computed on pairs of annotators for inter-annotator agreement.

[16] crowd-enVENT: `https://www.ims.uni-stuttgart.de/data/appraisalemotion`.
deISEAR and enISEAR: `https://www.ims.uni-stuttgart.de/data/deisear`.

coders, but also by first-hand emoters, whose judgments are too often absent in existing corpora. This allowed me to interrogate how well readers grasp the writers' intended emotional imports.

Of the three, deISEAR and enISEAR permit to ask that question in a multilingual setup: the former is a corpus of 1001 German texts annotated with seven emotion classes, the other is a companion English resource built analogously, to disentangle the effects of annotation setup and language when comparing to the original ISEAR resource. crowd-enVENT is similar but larger and with a focus on appraisal theories. For this reason, it can spur research directions beyond affective phenomena: it can support the more general study of human evaluations of real-life circumstances; it represents a valuable benchmark for investigation tracks interested in differences between annotation perspectives (generators vs. validators) and between event evaluation perspectives (cf. our analysis of mentioned entities); and psychology can regard it as a NLP counterpart of previous work, which reveals that a rich set of appraisal dimensions transfers in the domain of language for the decoding of internal states.

To conclude, I will underline the points that stood out from the study of these corpora, starting from a reflection on our study design, touching upon the benefits and lacunas of an appraisal-aware approach to emotions, and closing with remarks about the recognition of discrete emotions.

6.1 Lessons on the Two-Phase Crowdsourcing Strategy

Our two-phase experimental "scaffold" ended up being so effective that we used it in both data collection studies without any refinement. It provided:

- a setting for us to administer only implicit emotions to the readers, by removing words closely related to the prompting labels from the preceding phase (achieving this objective would not have been immediate with pre-existing texts);
- a way of gathering events while avoiding the hurdle of defining "event" (which would have been required with texts extracted from available sources);
- a highly structured setup where writers and readers faced two sides of the same task; in the appraisal-enriched scenario, they were guided to reason upon the same event properties.

This approach is open to variants that upscale our understanding of the appraisals role in emotion annotation. One research question for future work stems from the order of subtasks in our questionnaires of crowd-enVENT. Does rating appraisals before labeling emotions influence the latter type of judgment for the readers? Facing the evaluations first could drive them to form a more detailed representation of the emotion-inducing event, and possibly to make more accurate emotion inferences.

Another development regards the time at which emotion experiencers appraise the events that they recount. In an ideal experimental setting, appraisals are as recent as possible, since the memory of events further back in time might be altered. For us, adding the temporal distance between the task and the event as a constraint on the generation of descriptions would have been inefficient (i.e., it is hard to ensure that participants respect such a constraint) if not counter-productive (i.e., as an additional cognitive load for the crowdworkers). This research path can be explored by future studies working in lab with the participants' emotions at present.

One reason that pushed us to tap into a communication framework was to account for a natural emotion production–recognition sequence. But the framework we used is extremely simplified and falls short in reproducing the multiple phases that everyday interactions are made of. Therefore, an appealing avenue is one where generators and validators take turns. Writers could provide feedback to the readers for them to update their emotion and appraisal understanding. In practice, that could happen by incrementally refining their descriptions, stressing out the features of an event and adding details they deemed salient, until the text contains sufficient information for the emotion to be as unambiguous as possible. Making the phases dynamic like so would have been challenging in our solution based on crowdsourcing.

6.2 Lessons on Appraisals

There are texts in crowd-enVENT that emanate a "common knowledge" emotion connotation, while others require more elaborate interpretations (e.g., by focusing on different parts of the texts, different appraisals might fit a description). Overall, a consistent number of descriptions received validations that matched the original event assessments. In fact, the readers proved capable to be correct about appraisals even if they failed to recover the gold emotion labels (RQ2).

This independence between the correct appraisal reconstruction and the correct emotion labeling foresees future developments for computational emotion analysis. It suggests, for instance, that measures of inter-annotator agreement can be adapted towards an account of fundamentally similar text understandings: emotion disagreements that come hand in hand with high appraisal agreement could be weighted as less relevant.

Results purported the advantage of appraisal theories for emotion annotation. Collecting judgments about event properties gave us the opportunity to analyze what lies behind a particular emotion choice: they revealed in retrospect why the readers picked a certain emotion label (e.g., they appraised the described event in a specific way), they disclosed the evaluations underpinning their disagreements with the writers (e.g., each of them appraised an event differently) (RQ3). By that, I do not intend to tout the superiority of a cognitive model of emotions over others per se. Appraisals are not sheltered from criticism (Prinz, 2005). For instance, event evaluations are in principle unbounded, not limited to 21 criteria; it can also be doubted that an appraisal, or the group of appraisal variables as a whole, is always sufficient and/or necessary for an emotion to happen. The value that appraisals hold from a NLP perspective is that they are simultaneously interpretable, as are basic emotion classes, and dimensional, as are VA(D) ratings in affect-based models. Whether all dimensions apply to all events or topics is a question open to exploration; at first sight that is implausible, having seen that appraisal judgments have different patterns across different emotions.

Acknowledging if and when (and which) appraisals have a systematic effect on the predictions made by classifiers has a promising application as well: empathetic dialogue agents could grasp internal states better by asking users to clarify the relevant evaluation dimensions (e.g., “Did you feel responsible for the fact?”, “Could you foresee its consequences?”). Lastly, these can be adopted with more varied, not experimentally-induced types of texts – an initiative that has recently taken its first steps with texts extracted from social media (Stranisci et al., 2022). The full potential of appraisal information in emotion-laden data might flourish with spontaneously produced and longer pieces of texts, which give both human annotators and classifiers some context to picture the evaluation stage of an affective episode, ideally more than the event descriptions of crowd-enVENT did.

6.3 Lessons on Emotion Recognition

On emotions, readers agree with other readers to a comparable extent as they do with the writers. Computing agreement among observers or between observers and emoters does not alter the impression of their decoding abilities: consistency approximates correctness. Among all outcomes of the analysis of crowd-enVENT, this one supports that each individual's background and personal sensibility brings as much "noise" as the annotation perspective (if not, validators, who share the perspective of external coders, would have been more consistent than correct).

I believe that the conclusions drawn here apply beyond our observational scenario (e.g., the sentence completion task). To go after a view that warrants the viability of laboratory experimentation in other fields, "laboratory microeconomics are real live economic systems [... and allow] the methods, objectives and results to be interpreted and perhaps extended" (Smith, 1982). On the one hand, the texts of crowd-enVENT, deISEAR and enISEAR are inherently different from the messages sent by agents in a non-isolated environment, comprising actions and a plethora of reasons to communicate emotions to other agents, for example to influence their actions and beliefs. However, they enabled us to move within a manageable framework with parametrized conditions: all writers produced emotion-related texts with the same goal, all readers attempted to reconstruct emotions at a certain point in time with respect to a description. Besides, many pre-existing texts produced in a non-controlled setting are not less constrained. Those that populate social media, for example, have limits on the number of words and on certain sensible topics.

The inter-annotator agreement achieved between generators and validators promises that readers are reliable annotators (RQ1), for their judgments are not altogether incompatible. In some of the recounted experiences, one can appreciate a strong inter-subjective validity through a more prototypical connotation of the events or linguistic material that better conveys the emoter's reaction. This finding backs up that emotions are linked to rationality in at least two senses: in virtue of their reasonableness, which makes the writers' response to certain events comprehensible for an observer–reader or perceivable as conventional (i.e., "I might feel this way too under similar circumstances", "it's normal to feel this way" (De Sousa, 1990)); and in an intellectual sense, since cognitive evaluations of the qualities of an event and its implications and analytical inferences partake in the emotion process (Scherer, 2011) and decoding. At the same time, and irrespective of the texts' language, hardly did we find many readers agreeing on the ground truth. Solving this impasse is unfeasible, but our two-steps procedure permits to at least appreciate when the perceived emotions match or do not match the intended emotions. As an alternative, future work could study if wrong emotion judgments are considered valid by the writers themselves, by extending the corpus construction task to a multi-label scenario, where the writers indicate various emotion interpretations acceptable for their experiences.

Knowing that the readers are imperfect is important, because it helps sharpen the ultimate goal of computational emotion analysis. As Mohammad (2022) puts it: "it is impossible to capture the full emotional experience of a person [...]. A less ambitious goal is to infer some aspects of one's emotional state". To that the current chapter adds, in sum, that such a goal includes seeing emotion truths as nuanced truths. Although incorrect, the judgments that diverge from the writers' labels are often hard to pin down as implausible (cf. confusion between *sadness* and *anger*), and in that sense, the writers' answers might be the correct (intended) answers, but not the only ones that their texts imply.

Brief, reconstructed emotions are not only a function of the text, but both of the text and of extralinguistic factors Θ_i, in such a way that there exists no unique mapping from text to emotions, but there are as many mappings as there are annotators (i). An appealing way to make use of this remark is personalized emotion classification, to model texts together with other variables belonging to the subject who reads and judges them, such as background and past experiences. Another, that the next chapter explores, is to investigate how those Θ_is play a role in emotion recognition. I will do that with the multiple variables of crowd-enVENT that have not been fully analyzed here (e.g., *ethnicity*, *age*, which provide a stepping stone to study agreement), retaining two central lessons from the analysis of crowd-enVENT: it is possible to come to grips with (dis)agreement patterns, and annotation studies can ask for more (annotation layers) to understand more.

Chapter 4

Explaining Disagreements with Extralinguistic Factors

Continuing on the issue of subjectivity in emotion annotation, this chapter questions the role of extralinguistic factors in emotion recognition. It selects a taxonomy of variables that instantiate three sources of disagreement previously identified in NLP: stable characteristics of the annotators (e.g., age), contextual aspects (e.g., confidence, or the annotators' beliefs about their performance), and features of the texts (i.e., emotion intensity).

We tap on past work in psychology to select the taxonomy and formulate two hypotheses. First, individuals agree more if they have similar traits – which turns out true. Second, based on the idea that people recognize emotions while implicitly assessing the correctness of such judgment, we further expect them to produce more agreeing annotations when confident that they performed well in the task – confidence indeed approximates inter-annotator agreement, and it proves correlated to emotion intensity: perceiving stronger affect in text prompts annotators to more certain classification performances. To probe our assumptions, we rely on the annotation layers present in crowd-enVENT, and we collect judgments from scratch for a subset of the Corpus of Contemporary American English.

Highlights

To a small extent, judgments correlate to the coders' personal features.

Some apparently unsolvable disagreements are, in fact, systematic.

Confidence, intensity and agreements are entangled.

The coders can inform us that they expect to disagree.

1 The Curse of Disagreement

My analyses so far have shown that the labels assigned to a text by readers are often mismatching, but their diversity cannot simply be dismissed as an error of the coders. In certain cases, opposite judgments were equally tenable, justified by the underlying event evaluations. The fact that many reads stemmed from the same text rather reflects the openness-to-interpretation of emotions in language. Utterances do not have just one correct understanding, and this remark is not limited to my study, nor to event descriptions, nor to emotion analysis (Uma et al., 2021).

Valuing divergent annotations would thus be methodologically sound for (at least some areas of) NLP. Disagreements can be useful information to train automatic models (Jamison and Gurevych, 2015; Plank et al., 2014a; Fornaciari et al., 2021). At the same time, they collide with a time-honored desideratum for the computational study of language, namely, a consistent similarity among the annotators' decisions. High agreement indicates that the annotators have reached a comparable comprehension of the guidelines, as a signal of their reliability, which in turn validates the appropriateness of the coding scheme for the considered texts (Artstein and Poesio, 2008). Computational emotion analysis, instead, finds itself tolerating particularly low inter-annotator agreement scores (Alm et al., 2005; Melzi et al., 2014; Schuff et al., 2017, i.a.).

Some workarounds to mitigate inconsistencies among judgments have been proposed in the field (Volkova et al., 2010; Haider et al., 2020). They regard the formulation of the annotation guidelines. For instance, coders can be tasked to reconstruct the writers' emotion or to specify one they are *evoked* by the text (Buechel and Hahn, 2017b). But ultimately, emotions are subjective. People's personal characteristics, like their cultural background, are a built-in factor of disagreement in the task of emotion recognition. That is a morsel of commonsense knowledge. It emerged from the previous chapter, which concluded that the readers' task could be formalized as $\arg\max_{E\in\text{emotions}} p(\text{E} \mid \text{text}, \Theta_i)$, the sampling of an emotion E given not only the text they were shown but also some external parameters (i.e., Θ_i, encompassing potentially any set of factors that describe the annotator i, such as her physical state, her mood, her mental dispositions).

What is less explored is if this knowledge can contribute to understanding people's emotion recognition performance in text. Can we use the extralinguistic factors that we believe responsible of different judgments, to gain a better grip of disagreements? Plenty of variables converge in the task and, needless to say, it would be impossible to tackle all of them. How to systematically study, for example, the personal values and past experiences that cause the coders to associate a textual unit with an affective state but not another?

2 Quantifiable Factors of Disagreement

As a solution, one can identify quantifiable features of individuals. Ideally, these features should abstract over life experiences, establishing a common ground among multiple annotators (e.g., the feature "age", as it is safe to assume that being in a similar age range implies sharing some background, cultural references, and so on).

Regarding possible sources of annotation disagreements in NLP, Basile et al. (2021) listed three that interact with one another. First, differences in world perception. They concern the private sphere of opinions and judgments for which socio-demographics variables (e.g., age, gender, cultural background) have a margin of accountability. Second, the annotation context. A coder could answer the same question differently at different points in time, due for instance to slips of attention (Beigman Klebanov et al., 2008). Third, the stimulus characteristics (a linguistic addition to the other two extralinguistic elements), like lexical, syntactic and semantic features, all of which can affect the ambiguities of a text and the understanding of its meaning.

To fill in the three groups with quantifiable variables, psychology comes in handy, because it documents the link between variations in emotion perception and a number of factors. Such factors (i.e., those touched upon in Chapter 2, Section 1.3) are theoretically appealing. They could have an effect on emotion recognition in language, but not all of them have been investigated jointly with textual emotion stimuli. They are also easily accessible from a practical viewpoint, since they correspond to the annotation layers in crowd-enVENT, described earlier in the thesis. The three groups of factors can therefore be populated as follows.

Stable Factors.[1] Researchers in psychology determined that the ability to recognize emotions varies with age (Widen, 2013), gender (Hall and Matsumoto, 2004; Hampson et al., 2006) and personality traits (Terracciano et al., 2003; Matsumoto et al., 2000). Another finding, already explored in the previous chapter, is that computing agreement between observers and emoters can change the impression of the observers' ability, in comparison to only measuring the inter-observer consistency. That study (mentioned in Ekman, 1972, p. 242) was interested in the cultural effects on the task, to verify if Japanese and American annotators perceived facial expressions similarly. Culture was later found to underpin variations in emotion judgments (Jack et al., 2009; Nelson and Russell, 2013), and it is formalized in crowd-enVENT by the self-assigned *ethnicity* variable.

crowd-enVENT provides information about two further "categories of experience" that can divide or unite annotators. One is their *highest level of education*, and the other is the *event familiarity* that the validators felt towards the real-life experiences they judged. The first type of data was collected in the wake of the link between education and correct emotion responses, as testified in psychology, with more educated people being more likely to produce the expected labels (Trauffer et al., 2013; Mill et al., 2009; Wolfgang and Cohen, 1988). Scores of *event familiarity* were gathered with the rationale that having an experience similar to the one mentioned in the text might influence the emotion association for that text.

Contextual Factors. While the above factors account for stable differences, it is possible to select some that relate to the annotation

[1] By "stable", I mean features that last in a relatively long temporal surrounding of the annotation task, but are not immutable.

context, such as the emotion that the coders felt right before undergoing the task. This variable was included in crowd-enVENT following past findings on mood congruity, i.e., the match between the internal state of a person and the recognized emotion (Niedenthal et al., 2000).

Contextual causes of disagreement can also be attributed to people's confidence, or their degree of certainty in the annotation task. Humans (roughly) know how well they can "read" emotions in others (Realo et al., 2003). They judge affective phenomena (e.g., by observing faces) and the correctness of such judgments at the same time. In other words, annotators can evaluate their own confidence with respect to their labeling decisions. This hints at a possible relation between confidence and inter-annotator agreement also in text. One could expect the annotators to be discordant when they feel uncertain about their answers.

Linguistic Factors. Another aspect involved in emotion recognition is emotion intensity (Juslin and Laukka, 2001). This variable can be ascribed to the content of the stimulus, as the strength of the portrayed experience. It would be intuitive to think that emotions are more confidently recognized if they are expressed with stronger magnitude. Consider, for instance, the intensity of the following pairs of expressions: *"The teacher exploded"* > *"He snapped his annoyed temper"*, *"sadder"* > *"a bit sad"*, *"ecstasy"* > *"joy"*. Yet, in the sentence *"We had to cheer him up; later, he was off the ground"*, readers can choose which part of the text to attend to (the challenge that the speakers undertook – not intense, or the effect they had – intense) and be very confident about either choice. Similar counter-examples reveal that the link between the perception of emotion intensity in text and self-perceived confidence is opaque, and leaves room for exploration.

Goal of the Studies. A leading motive of the thesis is the cross-pollination of computational emotion analysis and psychology. I pull the threads between the two fields also in this chapter. My goal is to understand if quantifiable variables previously linked to emotion recognition from faces and audio stimuli (and a few more, such as *event familiarity*) also apply to written verbal data (in English). I analyze the role of factors pertaining to individuals, context and stimuli in the (dis)agreements that emerge when annotating emotions. Moreover, I brings into play event appraisals, like in Chapter 3.

The chapter poses three research questions. (RQ1) Are annotators of appraisals and emotions more reliable if they have particular properties? (RQ2) Are (dis)agreements with respect to the presence of emotions related to confidence? (RQ3) Are judgments of intensity and confidence entangled? The first research question has been partially answered in computational emotion analysis. While creating a VAD lexicon, Mohammad (2018) collected data about the annotators' age, gender, and personality traits, pointing out a significant relation between (nearly all) variables and inter-annotator agreement. However, that work did not dispose of the writers' ground truth, while crowd-enVENT does. Plus, it only examined a handful of factors, and whether its outcome can be meaningfully generalized to other types of annotations (discrete emotions and appraisals at the level of sentences) is up to debate. I fill in these gaps.

The judgments I will deal with are produced under conditions that supposedly spur different extents of disagreement (Buechel and Hahn, 2017b). One requires the readers to reconstruct the writers' affective reactions. The annotations of crowd-enVENT are of this type. The other scenario tasks readers to express their own perception of a text. Corpora annotated this way are available, but they lack information about the multiple factors of interest here. For that reason, I conduct an annotation study from scratch, labeling part of the Corpus of Contemporary American English (Davies, 2015) with emotion data and meta-data.

The two resources I use differ with respect to the type of texts, annotation perspective, and emotion labels they contain. I will consider different factors on each of them:

- To answer RQ1, I will use crowd-enVENT. Its structure lends itself well to understanding if agreement among readers changes according to their emotion state when performing the task (as a contingent factor) and to more stable individual differences (demographics, personality traits, etc.). Section 3 focuses on this question.

- I will answer RQ2 and RQ3 using the Corpus of Contemporary American English, by investigating the relationship between three types of judgments: about the presence of emotions, about their intensity (a stimulus characteristic), and about the confidence of the annotation decision (a contingent factor). Leveraging such information, I seek to grasp in what cases annotators differ or, on the contrary, concur regarding the verdict that an emotion is expressed in the text. RQ2 and RQ3 are addressed in Section 4.

We find that personal factors have a significant impact on agreement, but each of them to a different degree. Confidence turns out to explain systematic differences in the decisions of the annotators; so does intensity; and impressions about intensity and confidence prove to be correlated. Based on the latter results, I introduce a strategy to leverage confidence and intensity and alleviate inter-annotator inconsistencies.

3 An Analysis of Agreement through Demographics, Personality Traits, Event Familiarity and Current Emotion

When building crowd-enVENT (cf. Page 86), we collected a rich set of properties of the annotators, both in the text generation phase (which produced 6600 event descriptions) and during validation (where a subsample of texts were labeled by readers). These annotation strata were ignored in the analysis of agreement in Chapter 3. Now, we put into question how they relate to emotion and appraisal judgements.

Each survey started by asking participants to rate their current *emotion state* on a 1–5 scale for twelve emotions; it concluded with a block of questions related to the variables of *age*, *gender*, *ethnicity*, *highest level of education*, and *personality traits*, with which individuals informed us about their relatively stable characteristics. The contributors in the text generation stage could take more than one survey, but they were required to fill in the closing questions only once. For those who did so multiple times, we averaged the *personality traits* scores (personality traits were to be rated on a scale, therefore such answers could vary slightly from one submission to another). Participants in the validation phase also rated their *event familiarity* on a 5-point scale.

Below, I briefly describe the statistics of these properties in the corpus (Section 3.1), and then discuss inter-annotator agreement "conditioned" on them (Section 3.2).

3.1 Distribution of Variables in crowd-enVENT

Table 4.1 reports the distribution of variables in the generation and validation sides of crowd-enVENT. For the *emotion state* prior to par-

		Gener.	Valid.
Gender	Female	1639	710
	Gen. variant/non conf.	43	22
	Male	690	480
	Prefer not to answer	7	5
Education	Secondary Education	226	69
	High School	738	356
	Undergrad. Degree	975	527
	Graduate Degree	379	223
	Doctorate Degree	38	35
	No formal qualifications	9	5
	Not applicable	14	2
Ethnicity	Austral./New Zeal.	65	21
	North Asian	–	2
	South Asian	59	41
	East Asian	55	34
	Middle Eastern	21	7
	European	1247	808
	African	58	28
	North American	550	178
	South American	10	1
	Hispanic/Latino	59	17
	Indigenous	12	5
	Prefer not to answer	53	11
	Other	190	64

		Gener.	Valid.
Ev. Fam.	1	–	1789
	2	–	765
	3	–	773
	4	–	1039
	5	–	1634
Age	(median)	28	36
Current State	Anger	1.27	1.17
	Boredom	2.24	1.88
	Disgust	1.17	1.09
	Fear	1.29	1.19
	Guilt	1.34	1.23
	Joy	2.06	2.20
	Pride	1.69	1.92
	Relief	1.69	1.77
	Sadness	1.60	1.46
	Shame	1.27	1.18
	Surprise	1.27	1.3
	Trust	1.95	2.33
Pers. Traits	Openness	2.24	1.97
	Conscient.	2.32	2.60
	Extraversion	−0.58	−1.18
	Agreeabl.	1.80	2.17
	Emot. stab.	−0.04	0.75

Table 4.1: Distribution of annotations in the generation and validation phases of crowd-enVENT. Counts of participants across labels of gender (Gen. variant/non conf.: gender variant/non conforming), education, ethnicity (Austral./New Zeal.: Australian/New Zealander), and event familiarity. Median age, average values of current emotion state (rated on a 1–5 scale) and of personality traits ranging in [-3,3]. Conscent.: conscientiousness, Agreeabl.: agreeableness, Emot. stab.: emotional stability.

ticipation in our study, the highest average value is held by boredom (2.24) followed by joy (2.06), trust (1.95), pride (1.69) and relief (1.69). The lowest value is observed for disgust (1.17). Values of *emotion state* are similar in the two phases within each emotion label prompting the texts. Concerning *personality traits*, the participants reported high scores of conscientiousness (avg. 2.3/2.6 in the generation/validation phases), openness (2.3/2.0), and agreeableness (1.8/2.17). The majority of people who disclosed their *gender* were female (generation: 1639, validation: 710), followed by males (690, 480), and a handful identifying with gender variants (43, 22). Their *age* distribution has a median of 28 at generation time and 36 in the validation step. Most participants had a high school-equivalent degree (generation: 738, validation: 356), an undergraduate degree (975, 527), or a graduate degree (379, 223), and only a few did not have any formal qualification (9, 5). Moreover, most people identified as European (1247, 808) or North American (550, 178).

In the remainder of the section, I will refer to each variable as "factor" (e.g., *gender*) and to its values (e.g., female, male) as "properties".

3.2 Conditioned Inter-Annotator Agreement

We now compute the agreement among coders. Our goal is to understand if such a score is influenced by any of the contextual and stable properties that the readers chose to describe themselves (RQ1). The structure of the data allows to answer the first research question in two ways. One analyzes whether validators with a given property (e.g., familiar with the described event) recognize the correct emotions better than those who do not have it. Observing if they are the ones more in agreement with the writers would be a direct NLP counterpart of research in psychology about people's emotion recognition ability. The other direction considers whether readers who share a property (e.g., both are familiar with the event) agree with each other more than readers characterized by another property (e.g., they are unfamiliar with the event) or by different properties (e.g., one is familiar, the other is not). We are more interested in this second approach that leaves the writers aside. This permits us to focus on the coders who took the task of emotion decoding, and to see how consistent they are, irrespective of how correct. The analysis of factors as conditions of correct emotion reconstruction is in Appendix B, Section 2.

3.2.1 Method

To investigate each factor and property thereof separately, we extract several subsets of annotations from crowd-enVENT. We subsample them as follows: (1) we pair validators; (2) we extract pairs that have a property of interest; (3) we compute agreement on the texts annotated by people sharing or not sharing that property.

Step (1) consists in pairing the coders who labeled the same descriptions, which returns 12,000 validator–validator (V–V) pairs, that is, $\binom{5}{2} \cdot 1200$ pairs constructed among the 5 validators annotating a text, out of the 1200 texts that entered the validation phase.

Step (2) partitions the V–V pairs according to different criteria. We aim at extracting pairs in which the annotators share a property or not. The decision of whether the property in question holds depends on the factor it belongs to. Regarding *gender*, annotators can be both males, both females, or each of a different gender; for *age*, we focus on age differences, as greater or lower than 7 years; for *event familiarity*, we establish a cutoff point of 3 and establish that a score >3 indicates that an annotator is acquainted with the event. Thus, we extract three groups of pairs: those in which both annotators are familiar with the event, pairs in which both are unfamiliar, and pairs in which they differ in this regard. We filter the coders in the three types of pairs also for each property of the *personality traits* factor. For instance, regarding openness we ask: Did both validators turn out to be open (i.e., with a score for this trait greater than a certain threshold)?, Did they turn out to be not open?, Was only one open?

The three-group divisions allow to observe the impact of specific properties on the annotation task. In principle, we could carry on this procedure with *ethnicity*, *highest education level*, and *current emotion state*, for instance by creating V–V pairs with a doctoral degree, pairs in which neither V has that degree, and pairs in which it characterizes only one annotator. However, compared to the other factors that allowed for a limited number of answers (e.g., only 4 values for *gender*), the annotations of *ethnicity*, *highest education level*, and *current emotion state* distribute across many more properties (e.g., 13 for *ethnicity*). Noticing that they hamper the creation of three substantial groups of pairs to compute agreement, we take a more coarse-grained approach with them. We consider if the annotators chose the same answer, irrespective of what that is, and we create two groups: one where each V in a pair has the same property (e.g., collapsing V–Vs who hold a doctoral degree with V–Vs who do not hold it), and one in which each V has a different property. This still permits to see if sharing a value for a factor impacts agreement.

Therefore, step (2) returns various subsets of annotated texts. Next, with step (3) we take the intersection between all subsets within a factor, (e.g., female–female, male–male, female–male for the factor *gender*) thus obtaining groups composed by pairs who annotated the same texts.

Using these, we compute agreement on each subset through multiple measures, as we did in Chapter 3 (Page 105). For emotions, we use average F_1 and accuracy. For appraisal annotations, we employ average RMSE scores.

3.2.2 Results

Table 4.2 details the results for the factors *gender*, *event familiarity*, *age* difference and *personality traits*. The first row (Pair type: "All data") reports the numbers obtained on the entire annotation set, already seen in Chapter 3. The other rows correspond to results obtained on subsets of texts extracted by factor (examples of agreements and disagreements of V–V pairs sharing or not sharing some selected properties are in Table 4.3). Note that the number of pairs on which agreement is computed (column "# Pairs") varies depending on the property under consideration, as different properties might hold for different numbers of people. The inter-annotator agreement scores within each row were obtained from the same textual instances. Therefore, for a given measure, we can compare the within-factor numbers, i.e., those inside a colored box, but not across factors. We calculate the statistical significance of their difference under a .95 confidence level via bootstrap resampling (1000 samples) (Canty and Ripley, 2021; Davison and Hinkley, 1997). Pairs of asterisks or dagger symbols indicate pairs of numbers within a box that are significantly different. All of them are, if three numbers are marked with a "*".

Concerning *gender*, we consider the groups with the most common self-reported answers (i.e., male, female). Across emotion measures, female–female pairs agree 7pp F_1 more than the male–male subset and 4pp F_1 more than the mixed-gender subset. *Event familiarity* (rated on a 1–5 scale) leads to significant differences in the appraisal assessments. On emotions, validators unfamiliar with an event agree slightly more compared to those who selected a strong event familiarity. A possible explanation is that readers who did not experience an event similar to the description relied more on the information emerging from the text than on their past knowledge.

To evaluate the impact of *age* on agreement, we separate the pairs at a threshold of 7 years (we tested other thresholds, which lead to smaller differences). All differences are comparably small (<3pp in F_1 and Acc.), but still significant for emotions.

			Agreement		
			Emotion		Appraisal
Factor	Pair type	#Pairs	F_1	Acc.	RMSE
None	All data	12000	.49	.52	1.48
Gender	male–male	1113	*.45	*.49	1.50
	female–female	1377	*.52	*.55	*1.50
	male–female	3772	*.48	*.52	*1.49
Event Fam.	familiar–familiar	540	.44	.47	*1.42
	unfamiliar–unfamiliar	676	.45	.48	*1.47
	familiar–unfamiliar	2445	.45	.48	*1.54
Age diff.	≤ 7	3839	*.51	*.54	1.48
	> 7	7991	*.48	*.51	1.48
Pers. Traits	open–open	1472	.49	.52	1.47
	not open–not open	1568	.48	.51	1.48
	open–not open	4560	.48	.51	1.48
	conscientious–conscientious	1638	.51	.53	*1.49
	not conscient.–not conscient.	1426	*.51	*.54	$^{\dagger *}$1.46
	conscientious–not conscient.	4596	*.49	*.51	†1.49
	extravert–extravert	1685	*.48	*.51	*1.51
	not extravert–not extravert	1535	*.52	$^{\dagger *}$.55	*1.46
	extravert–not extravert	4830	.50	†.52	*1.48
	agreeable–agreeable	1451	*.51	*.54	1.47
	not agreeable–not agreeable	1553	$^{\dagger *}$.45	$^{\dagger *}$.49	1.47
	agreeable–not agreeable	4506	†.49	†.52	1.48
	emo. stable–emo. stable	1621	.49	.52	$^{\dagger *}$1.50
	emo. unstable–emo. unstable	1559	.51	.55	*1.47
	emo. stable–emo. unstable	4770	.50	.53	†1.48

Table 4.2: *Gender, Event Familiarity, Age* and *Personality Traits* as conditions of validators' agreement, shown as F_1 and Accuracy (for emotions) and average root mean square error (for appraisals). For each factor, the column "#Pairs" reports the size of a sample on which agreement is computed. Colored boxes indicate numbers that can be compared to each other. "*" and "†" indicate that they are significantly different, as found with 1000× bootstrap resampling, confidence level .95.

	Pair type	Emotions	Example
Gender	male–male	guilt–guilt	I dropped a light bulb at the department store and it broke, then I put it back on the shelf and got a new one.
	male–female	guilt–shame	
	female–female	anger–anger	I was wrongly accused of stealing
	male–female	anger–surpr.	
Age diff.	> 7	bored.–bored.	Endlessly doomscrolling through social media, then I think to myself ”I need to do something”
	≤ 7	bored.–guilt	
	>7	sad.–guilt	I thought I had upset a fiend and damaged our relationship
	≤ 7	sad.–sad.	
Event Famil.	> 3	joy–joy	I was dancing with my partner and closest friends right before the pandemic
	≤ 3	joy–fear	
	>3	trust–relief	My friend needed support to manage her money so she gave me control of her bank accounts
	≤ 3	trust–trust	

Table 4.3: Examples on which all V–V pairs judging a given text and sharing a property agreed on the assigned emotion labels. On the same texts, V–V pairs who differed by that property disagreed. For instance, all male–male pairs agreed on the emotion *guilt*, and on the same text, all mixed-gender pairs disagreed; similarly, for the third description, pairs of validators characterized by Age diff.$>$7 agreed on the label *boredom*, while in pairs with Age diff.$\leq$ 7 one validator picked *guilt*.

For the analysis of the influence of validators' *personality traits*, we split the validators with a threshold that approximates a balanced separation of all judges. While openness and emotional stability have no significant impact on the emotion agreement measures, agreeableness does, with agreeable–agreeable pairs reaching significantly higher F_1 and accuracy scores than the non agreeable and mixed counterparts. Further, non extravert pairs have a significant improvement over the extravert V–Vs across annotation variables and measures. Pairs with a low level of conscientiousness are significantly more in agreement than conscientious–not conscientious pairs.

			Agreement		
			Emotion		Appraisal
Factor	Pair type	#Pairs	F_1	Acc.	RMSE
Education	=	3589	.50	.53	*1.47
	≠	8231	.49	.52	*1.48
Ethnicity	=	4022	.49	.52	*1.47
	≠	6238	.48	.51	*1.50
Current State	anger =	1036	.50	.53	*1.47
	anger ≠	724	.51	.54	*1.53
	boredom =	4648	.49	.52	*1.47
	boredom ≠	4472	.49	.52	*1.49
	disgust =	660	*.51	*.53	*1.46
	disgust ≠	480	*.42	*.45	*1.61
	pride =	4784	.50	.53	*1.47
	pride ≠	4916	.49	.52	*1.50
	relief =	4120	.50	.53	*1.47
	relief ≠	3920	.48	.51	*1.52
	sadness =	3040	*.48	*.52	*1.48
	sadness ≠	2380	*.51	*.54	*1.46
	shame =	1258	*.45	.48	1.47
	shame ≠	862	*.48	.51	1.49
	surprise =	2060	*.53	*.55	*1.47
	surprise ≠	1610	*.46	*.49	*1.56
	trust =	5324	.50	*.53	1.48
	trust ≠	5916	.49	*.51	1.48

Table 4.4: *Education*, *Ethnicity*, and *Current Emotion State* as conditions of validators' agreement, shown as F_1 and Accuracy (for emotions) and average root mean square error (for appraisals). For each factor, the column "#Pairs" reports the size of a sample on which agreement is computed. Colored boxes indicate numbers that can be compared to each other. "*" indicates that they are significantly different, as found with 1000× bootstrap resampling, confidence level .95. *Current Emotion State* only reports emotions for which we found significant differences.

	Pair type	Emotions	Example
Emotion State	disgust =	anger–anger	my partner gaslights me to get money from me for alcohol and cannabis
	disgust ≠	anger–disgust	
	sadness =	relief–relief	Patient accepted elaborated treatment plan easily
	sadness =	relief–fear	
	shame =	fear–fear	we were attacked
	shame ≠	fear–surpr.	
	surprise =	trust–trust	I form a meaningful connection with someone that shares my beliefs, listens to my thoughts and communicates with me in a respectful way
	surprise ≠	trust–joy	
	trust =	guilt–guilt	I accidentally broke one of my mother's cats legs
	trust ≠	guilt–sad.	

Table 4.5: Examples on which all V–V pairs agreed on the emotion label of a text (column "Emotions") while sharing an emotion state ("=" indicates that both V–V rated the emotion state as above or equal to 3, or below this threshold). On these texts, V–V pairs who differed by that state ("≠") disagreed. For instance, all pairs in the same sadness-related state agreed on the emotion *anger*; in mixed-state pairs, one annotator disagrees by choosing the label *surprise*.

Results for the factors *education, ethnicity*, and *current emotion state* are in Table 4.4. In the column "Pair type", "=" indicates that both judges picked the same answer for a given factor. Vice versa for "≠". In the rows corresponding to *current state*, "=" signifies that both coders in a V–V pair rated a particular emotion as above or equal to 3, or below this threshold; "≠" marks cases in which one coder was above and another below it. We do not report emotions where no significant difference was found (i.e., *fear*, *guilt*, and *joy*).

Pairs sharing the same education level and self-assigning the same ethnicity turn out to have a higher shared understanding of emotions and appraisals than mixed pairs. However, only for appraisals the difference is significant. For *current state* results are diverse. Except for *shame* and *trust*, all emotions lead to significant differences in the appraisal agreement. In general, people sharing an emotion state achieve better RMSE scores than the "$\neq$" groups. The only case in which the opposite happens is *sadness*. Lastly, considering F_1 and accuracy scores obtained on the annotated emotions, the current states of *disgust*, *surprise* and *trust* show higher in-group agreement (accuracy) when the annotators are close in regard to these three states. Table 4.5 reports example texts corresponding to the statistically significant results in the column "Emotion" of Table 4.4.

Interpreting these results is not equally straightforward for all factors, because in some of them, namely *personality traits* and *current emotion state*, the validators' answers are not mutually exclusive, in others they are. Moreover, the numbers just seen raise an important question: How to explain a property's positive impact on emotion which does not mirror an impact on appraisals (or vice versa)? The reasons behind such a difference go beyond the scope of this chapter, but the significant trends we have found are sufficient to answer our first research question. The analysis of inter-annotator agreement conditioned on self-reported personal information revealed that better agreement on emotions and appraisals is favored by the possession of certain properties[2] (e.g., females agree more than males), as well as by the sharing of certain properties (e.g., experiencing a state of *trust* has a positive impact on agreement compared to not sharing it). Our results based on discrete and appraisals models on emotions echo those obtained by Mohammad (2018) in a VAD framework. However, differences between groups of judges with diverse features are small (all F_1 and accuracy scores overall have the same order of magnitude), and they vary depending on the property of a factor that one considers (e.g., *trust* vs. *sadness*).

4 An Analysis of Agreement through Emotion Intensity and Annotators' Confidence

The above analysis focused on emotion and appraisal recognition as a task to reconstruct what others felt, but it is equally important to understand the (dis)agreements that emerge with judgments that account

[2] In Appendix B, Section 2, those properties are shown to affect also the reconstruction of the *correct* emotion labels and appraisal ratings. That is, they promote both better agreement (observed here among V–V) pairs and better emotion recognition performance (observed in Appendix B with an analysis of agreement among G–V pairs).

for the coders' emotional point of view. This is a type of information relevant for the field, and it is what I consider in the remainder of the chapter.

I proceed to study only two factors. One is intensity, pertaining to the perception of verbal stimuli, which is a dimension of interest for many studies in computational emotion analysis (Strapparava and Mihalcea, 2007; Aman and Szpakowicz, 2007, i.a.). The other is confidence, a circumstantial element that I investigate with emotions, but which could be easily quantified also with other phenomena, as a self-reported degree of the coders' (un)certainty about their annotations. The collection of this kind of judgments is not a common practice. Focusing on emotions, past research has found that self-assigned scores of confidence are predictable based on some vocal attributes of emotion speech stimuli (Lausen and Hammerschmidt, 2020), but this has never been done on written material, to the best of our knowledge.

We bring the two factors together in an annotation study based on emotion recognition, in which self-perceived confidence and emotion intensity are dimensions to be rated. Given a subset of the Corpus of Contemporary American English (COCA), raters distinguish emotion-bearing sentences from neutral ones, while quantifying both the intensity of the emotion and their confidence[3] on a Likert scale (in line with Bègue et al., 2018). This annotation differs from the one conducted on crowd-enVENT with respect to many parameters:

- Data: we use pre-existing texts, which span over implicit and explicit emotion (as well as non-affective) expressions. Such texts are not restricted to event descriptions.
- Emotions: as we annotate a small sample of texts, we simplify the task to a binary setting, distinguishing *neutral* sentences from *emotion*-bearing ones. This allows us to obtain a consistent number of items per label.
- The circumstantial factor: as opposed to the current *emotion state* used earlier, confidence can change from the annotation of a textual unit to another (in this sense, it interacts with the stimuli features).
- Annotation coverage per coder: all annotators label all textual instances, which allows for a more straightforward computation of agreement among them.

[3] Also crowd-enVENT includes the dimensions of intensity and confidence, but these were collected with an emotion decoding task that differs from the current one.

- Annotation perspective: we collect annotations without tasking the readers to infer the writers' states, but rather asking about their personal associations with the text, which we expect to spur more extreme disagreements.

This last design decision defines an experimental condition more similar to past research (Strapparava and Mihalcea, 2008; Bostan et al., 2020; Haider et al., 2020). That is ideal to show that studies can better understand emotion disagreements, even by considering as little additional information as intensity and confidence (collecting annotators-related bits of information like those we used in crowd-enVENT, and doing so for thousands of datapoints, is laborious and not always feasible).

If confidence has to do with (dis)agreements (RQ2), it may serve as a diagnostic tool for systematic differences in annotations. Moreover, at first sight confidence seems entangled with intensity (RQ3). If that can be confirmed, such a link becomes relevant for annotation and modeling studies, because it opens the question wether automatic regressors that predict intensity actually model this variable, or rather capture the coders' self-perceived confidence.

4.1 Collecting Emotion Judgments from the Readers' Perspective

The first step in this study is to collect emotion assessments on diverse types of texts. Below is a description of the annotation task we instantiate and the data we use.

Task. We are not interested in which emotion people interpret from text, but rather if they recognize any. Judges read sentences and answer the question *(*EMO*) Is it Emotional or Neutral?* We instruct them to consider the presence of an emotion only with respect to their personal viewpoint. For the items deemed to express an emotion, we also ask *(*INT*) How strong is it?*, which enables us to obtain ratings about affective strengths on a Likert scale from 1 ("not intense") to 3 ("very"). Lastly, we have raters self-evaluate their own judgments on a scale from 1 ("unsure") to 3 ("certain"), in response to the question *(*CONF*) How confident are you about your answer to* EMO*?* The exact annotation guidelines are in Appendix A, Section 2.

We opt for an in-lab setting. Raters are three female master students aged between 24 and 27, proficient in English, and with some annotation experience and background in computational emotion analysis.

	P	R	F_1
Emotion	.88	.84	.86
Neutral	.89	.92	.90

Table 4.6: Binary (automatic) classification tested on the UED test set.

Data. Corpora that include emotion classes are usually tailored on specific domains. We broaden our focus to multiple genres, and annotate sentences from the 2020 version of COCA.[4] COCA was not curated for emotion analysis. It includes unlabeled texts that occurred from 1990 to present in different domains, like blogs, magazines, newspapers, academic texts, spoken interactions, fictions, TV and movie subtitles, and web pages.

With a corpus of this size (>1B words), considering all data points would be costly, and randomly selecting them could cause imbalance in the final annotation – i.e., a majority of *neutral* instances. Therefore, we draw a sample biased towards emotional sentences with a combination of rules and classifier-based information. To obtain such a classifier, we fine-tune the pre-trained BERT (Devlin et al., 2019) base-case model[5] on a number of emotion analysis resources, adding a classification layer that outputs the labels *emotion* or *neutral*. Data are the resources by Liu et al. (2007), Alm et al. (2005), Li et al. (2017), Ghazi et al. (2015), Mohammad and Bravo-Marquez (2017), Mohammad (2012) and Schuff et al. (2017), plus enISEAR (cf. Chapter 3) and ISEAR (Scherer and Wallbott, 1997). We make their format homogeneous with the tool made available by Bostan and Klinger (2018); next, as the labels in the resulting unified emotion data (UED) are not binary, we map the *neutral* and *no emotion* instances into *neutral*, and the rest into *emotion*. The total 136,891 sentences are then split into train (70%), validation (10%) and test (20%) sets. The classifier's performance on the UED is in Table 4.6. Given the similarity of domains between UED and COCA, we expect these numbers to be representative of the use of the classifier on our unlabeled texts.

Having that, we filter academic texts out of COCA for their arguably impartial language, and from each of the other genres, we randomly

[4] `https://www.english-corpora.org/coca/`.

[5] Using HugginFace: `https://huggingface.co/transformers/`.

	IAA
A1–A2	.38
A2–A3	.43
A3–A1	.30

(a) Cohen's κ for annotator pairs on EMO.

	Counts	
	1 vs. 2	3 vs. 0
E	138	304
N	170	88

(b) Counts of *emotion* (E) and *neutral* (N) items for EMO answers, aggregated by agreement (1 vs. 2, 3 vs. 0).

	CONF	INT
1	−.001	.04
2	.03	.20
3	.39	.30

(c) Fleiss' κ on EMO for each value of CONF and INT.

Table 4.7: Inter-Annotator Agreement on COCA.

pick 500 sentences, excluding those containing words that are masked for copyright reasons. Out of these, we sample 100 sentences balanced by class, i.e., 50 labeled as *neutral* by our classifier, 50 determined to bear an *emotion*. Thus, the annotators are shown 700 items, 100 per domain, with a balanced class distribution according to the classifier.

4.2 Inter-Annotator Agreement

We can now address RQ2 (Are (dis)agreements with respect to the presence of emotions related to confidence?) and RQ3 (Are judgments of intensity and confidence entangled?). As opposed to the study conducted on crowd-enVENT, all textual instances are annotated by all coders. Hence, we compute agreement using typical measures, such as Cohen's κ (1960) and Fleiss' κ (1971).

4.2.1 Results

We start by observing the annotators' agreement on EMO. As reported in Table 4.7 (a), the highest Cohen's κ between pairs of human judges is .43; Fleiss' κ for the three annotators is .34. These numbers appear unsatisfactory. They are partly due to the skewed class distribution in the annotators' choices[6], but they can also be traced back to how the EMO task was formulated: asking if a text is emotional from the readers' point of view (i.e., it "describes an event [...] to which *you*

[6] With skewed class distribution, chance agreement increases, penalizing the resulting κ (Cicchetti and Feinstein, 1990).

would associate an emotion", see Appendix A, Section 2) paves the way for heterogeneous responses, as we assumed.

A look at other IAA measures, like the absolute counts of items that are assigned to each label, leads to a more detailed picture. Table 4.7 (b) breaks down the annotated categories by agreement: column "1 vs. 2" corresponds to the groups of items on which one annotator chose a label, while the majority opted for the other; column "3 vs. 0" shows how many times all three annotators agreed. We see that 138 sentences out of 700 were deemed emotionally-charged by only one person (and hence, were associated to *neutral* by the other two). Two annotators picked the *emotion* class for 170 sentences, i.e., those which were *neutral* according to just one rater. As the amount of considered judgments increases, so does the inter-subjective validity of *emotion*. This tendency is clear in column "3 vs. 0", which shows that there were more emotional instances with three identical labels than those with conflicting ratings: perfect agreement was reached for 392 items (304 *emotion* and 88 *neutral*), more than half of the data, suggesting that people reached a certain degree of shared emotional understanding from the texts.

Confidence Approximates Disagreements (and so does Intensity). We now observe disagreements more closely. Table 4.8 reports the distribution of the confidence and intensity scores for the items where pairs of annotators picked different emotion labels. A row comprises all items on which either annotator (i.e., the one on the corresponding column) chose the *emotion* label and the other selected the *neutral* one.

Annotators seem to have had divergent judgments on the presence of emotions in a systematic way. Their inconsistencies correspond to certain patterns in the ratings of two factors under consideration. What emerges from the table is that the coders rarely disagreed when the *emotion*-leaning annotator had extreme confidence or perceived very high intensity. For instance, A1–A2 disagreed in total 201 times: on 24 sentences, A1 made the *emotion* choice, and on 177 sentences A2 picked the class *emotion*. Out of the 24 items, A1 rated 23 as having low intensity and 1 as medium intensity; out of the 177 sentences, 149 are considered of low intensity by A2, 27 as mild, and only 1 as highly intense. On 11 sentences annotator A3 made the *neutral* choice, while annotator A1 picked *emotion*, but rated such items with confidence 1 (5 sentences), or 2 (6 sentences) – never using the highest degree of confidence. The same holds for intensity: A1 rated 10 of the 11 sentences as having the lowest intensity, and 1 sentence as having

	A1						A2						A3					
	Conf			Int			Conf			Int			Conf			Int		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
A1-A2	15	8	1	23	1	0	54	95	28	149	27	1	–	–	–	–	–	–
A2-A3	–	–	–	–	–	–	13	25	6	37	7	0	56	61	3	94	26	0
A3-A1	5	6	0	10	1	0	–	–	–	–	–	–	95	128	17	169	70	1

Table 4.8: Distribution of Int and Conf for disagreements in the Emo task. For a pair of annotators (rows), disagreements are counted when one annotator (i.e., on the columns) chooses the *emotion* class and the other the *neutral* class.

intensity 2 – none with the highest intensity value.

These results answer RQ2: the evaluation of intensity and the self-evaluation of confidence underlie disagreements in discrete emotion annotations. Further, they corroborate that different intuitions are often not totally incompatible, since within the disagreeing pairs, the annotator who took the *emotion* decision did so without being extremely confident.

We strengthen this finding by looking once more at Fleiss' κ, as a standard measure of IAA. We compute it for the answers of Emo, like we did before. However, this time we consider to be emotional only the items on which all annotators who picked the label *emotion* also chose a certain level of confidence or intensity. Table 4.7 (c) displays κ separately for different levels (rows) of Conf and Int. These values are still low, but they inform us that κ is affected by both factors. With lower Conf/Int, disagreements become more prominent. Vice versa, the highest IAA (.39) is achieved on the most confident answers.

Stronger Intensity, Higher Confidence. Having found that confidence and intensity have a direct relationship with disagreements, we analyze how they link to one another. We focus on the ratings of the 304 sentences with the unanimous *emotion* judgment. Using these, we compute the intra-annotator correlation between the answers of Int and the corresponding ratings of Conf. Spearman's ρ (Spearman, 1904) reveals a significant positive correlation between intensity and confidence for all annotators (ρ = .5 for coder A1, .58 for A2, and .64 for A3, p-value $<$.05 for all). This answers RQ3, as it suggests that the three readers believed they correctly classified a text if they also perceived a high emotion intensity.

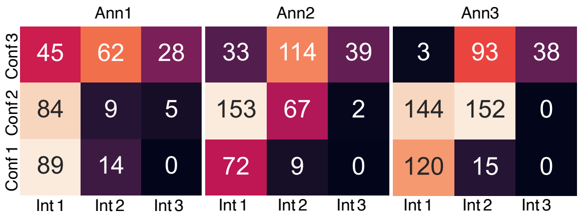


Figure 4.1: Cross-tabulation of Int and Conf by annotator, for the items that each of them deemed emotional (darker colors indicate fewer intems).

Figure 4.1 gives an in-depth account of the Conf-Int relation. It plots the counts of items labeled with a certain emotion intensity and a certain confidence level, separately for each annotator. The columns "Int3" tell us that the annotators rarely perceived intensity as strong without being extremely sure that the text expressed an emotion. In fact, no instance received the highest intensity and the lowest confidence (Int3-Conf1) at the same time. Conversely, for cases of low intensity, annotators tended to stay low also on the scale of confidence.

On What do People Agree? We gain additional insight into the readers' judgments with a manual analysis of the data. Perfect agreement occurs in the presence of certain textual patterns. Items unanimously considered emotional often report personal impressions about a state of affairs, or about the speakers' interlocutor (e.g., "*Paris is so sexy*", "*Your expression changed from excited puppy to crestfallen*"). Mostly, they depict first-hand experiences of the speakers (e.g., "*We'll miss you, but we'll be watching*", "*I'm afraid I don't see anything very beautiful right now*", "*Others helped me and it made a huge difference*"). Instead, sentences that received three *neutral* labels are more centered on objective statements, like "*Furthermore, the types of materials of manufacture are different*", and "*They continue walking*".

One difference between the *emotion* and *neutral* labels is the frequency with which annotators agreed on either. We find that they concurred more on the former. This invalidates our expectation that, not being given a varied set of affective categories, and not identifying *what* emotion they are judging, people would have resorted to the neutral choice. Moreover, annotators converged more on one or the other label depending on the genre of a text: looking at the distribution of the 304 unanimous *emotions* (magazine: 28 sentences, blogs: 44, news: 27, tv: 67, fiction: 54, spoken: 39, web: 45) and the 88 *neutrals* (magazine: 18

Conf	Int	Text
1	1	I was always a little wary of Arya and Sansa (who also did a little Stoneheart-style vengeance last year) taking on their mother s role .
1	1	You can't stress because you just have no idea what 's going to happen.
1	2	Mmm , Lordy , Lordy , Lord have mercy .
1	2	Frost is trying to reconcile impulse with a conscience that needs goals and harbors deep regrets.
2	1	" I'm fine ," he replies absently , eyeing the open book .
2	1	My sister likes her map :) HI Chery :) lol I'll take'em where I can get'em ...
2	2	The soldier talks about child detainees .
2	2	We did n't get to bury the others .
3	1	I bruised my lip .
3	1	"Chalk is unforgiving, " says Oates .
3	2	They looked happy, confident .
3	2	I hope I can work through my feelings and keep his friendship in my life.
3	3	I will completely destroy them and make them an object of horror and scorn , and an everlasting ruin.
3	3	If – if I could die and bring her back , I would , but I can't , and I have to deal with that now.

Table 4.9: Sentences on which the annotators reached perfect agreement on Emo, Conf, and Int. All texts were assigned the label *emotion*.

sentences, blogs: 12, news: 22, tv: 4, fiction: 5, spoken: 12, web: 15), we see that people recognized that affect often manifests itself in fictions, for instance, but is rarer in news – the opposite holds for the *neutral* expressions.

An obvious strategy to recognize emotions could have been to find an explicit emotion word in the text. But this was not the case. Explicit expressions were actually considered less emotionally intense than others: the majority of sentences with Conf3-Int2 contain emotion words (e.g.,*"I was sad to leave."*), while those with CONF3-INT3 are related to extremely negative states of mind conveyed through implicit expressions (e.g.,*"[...] if I could die and bring her back , I would , but I can't , and I have to deal with that now"*).

Another noticeable trend is that the annotators who gave the same score to intensity (i.e., perfect agreement is both on EMO and INT) tended to agree on confidence as well (i.e., there are always at least two people with the same CONF), and vice versa. This could be a reflection of the correlation between the two variables.

Moreover, some items were scored with perfect agreement on all questions. This is the case for the sentence *"I can't believe that you saved my life"*, considered to convey an emotion of intensity 3 with the highest level of confidence. Similarly, *"Get off my back!"* was deemed to have a mild intensity, and *"'You have such an interesting life', she said, after a little small talk"* of intensity 1. While these examples were all rated as highly certain (CONF3), Table 4.9 shows some sentences on which the annotators reached perfect agreement across different confidence-intensity combinations, having chosen the label *emotion*. Note that there is no instance that elicited a high intensity evaluation (3) with medium or low confidence (either 2 or 1). The opposite cases (e.g., CONF3 and INT1), instead, are present in the data, indicating that it is harder to be uncertain of a strong emotion than to be sure of a weak one. They also provide qualitative evidence that the correlation between confidence and intensity, though intuitive, has counterexamples. Lastly, the table suggests that our readers had greater certainty with texts that leave less room for ambiguity in interpreting the author's emotional state or intentions. In comparison to the texts with CONF1 and INT1, those marked as CONF3 and INT3 evoke concepts that are more straightforwardly emotional (e.g., destroying, facing an irreversible misfortune), and are conveyed directly, through impactful language (e.g., via intensifiers like "completely"). In such texts, empathy with the writer might have played a role in reducing the annotators' uncertainty.

4.2.2 Post-Processing Disagreements

If systematic differences among annotators can be diagnosed with the help of confidence and intensity, can they also be resolved to some extent? For our final analysis, we use the CONF and INT scores as acceptance thresholds for the label *emotion*, to post-process the EMO decisions of each judge: they turn into *neutral* if the corresponding

	Emo Conf<2		Emo Conf<3	
	1 vs. 2	3 vs. 0	1 vs. 2	3 vs. 0
E	172	187	141	56
N	169	172	73	430
	Emo Int<2		**Emo Int<3**	
	1 vs. 2	3 vs. 0	1 vs. 2	3 vs. 0
E	165	82	57	7
N	118	335	17	619

Table 4.10: Counts of labels for subsets of ratings on Emo, post-processed with acceptance thresholds <2 and <3, for both Conf (top) and Int (bottom).

Conf or Int answer does not reach a certain threshold t. For instance, using Int as threshold, with t<2 all items labeled *emotion* in Emo are kept as such only if their Int is 2 or more, all others are mapped to *neutral*.

Agreement counts on the post-processed annotation of Emo are in Table 4.10. We see that the number of agreed upon items increases by increasing the sets of equal ratings, similar to what we observed in Table 4.7 (b). For instance, 283 sentences received two unanimous judgments (column 1 vs. 2, under Emo Int<2), and 417 received three. In comparison to the original annotations broken down in Table 4.7 (b), we can observe a considerable change in the number of items with perfect agreement. While in the raw judgments they were 392, with t<3 they increase to 626. We find a similar pattern when leveraging confidence: with t<2 (low confidence), it is obtained for 359 items, and with t<3 (moderate), perfect agreement increases to 486 items.

This comes at the cost of agreeing on fewer *emotion* sentences (304 before filtering, 7 and 56 after applying the highest threshold to Int and Conf), but it indicates that the better raters agree on intensity or confidence, the more they agree regarding the presence or absence of emotions.

5 Discussion and Conclusion

The two analyses I presented have been driven by the idea that disagreements are not necessarily symptomatic of unreliability. This claim is espoused by a whole body of research in NLP, committed to understanding the ways in which human intuitions should be aggregated and evaluated, and the reasons and the patterns of their disagreement (Bayerl and Paul, 2011; Bhardwaj et al., 2010; Qing et al., 2014; Peldszus and Stede, 2013; Plank et al., 2014b; Sommerauer et al., 2020, i.a.).

Discussions of this type have not received great attention from emotion annotation studies so far, but their applicability to emotions should not come as unexpected. Emotion assessments derive from many factors. Therefore, knowledge of such factors can enrich and eventually explain affective judgments, when these alone are not sufficient to measure the quality of textual annotations. Such knowledge can further inform studies dealing with automatic emotion modeling, which might proceed by recognizing the importance of the spectrum of perspectives (and the underlying extra-linguistic correlates) that people have about emotions in text. In fact, while the current chapter concentrates on the human side of emotion analysis, it fundamentally advocates finding methods that determine if, when, and why multiple automatic decisions can coexist for the same text, with the final goal of achieving persona-specific automatic emotion recognition – a modeling setup where a system adopts the viewpoint of an individual with specific characteristics, as opposed to relying on a "gold", average emotion interpretation.

Below, I provide a critical recap of the main findings for the three groups of variables investigated in the chapter, and I point out how to advance work on emotion disagreements.

Stable Factors. We studied inter-annotator agreement on fine-grained emotions and appraisals by considering if the readers of crowd-enVENT had particular characteristics. The validators' features and their reliability were found to have a significant relationship, but differences between the annotations of people with various features were small (RQ1).

Knowing that the properties of individuals did not have a striking effect on their performance is encouraging, in a way. It excludes a series of expensive-to-measure factors as prominent sources of disagreement. We could have anticipated that. Even if a group of people has a defined socio-cultural influence with specific life experiences or values, there is no standardized way of perceiving emotions. As Basile et al. (2021) notices, no variable is sufficiently powerful to capture the uniqueness of each coder.

Still, the results we discussed raise important items for future research. For one thing, we need to understand if they are an artefact of the events in crowd-enVENT. The circumstances recounted in it do not display, for instance, a strong culture specificity that could reflect in *ethnicity* "biasing" the annotators' emotion associations; similarly, one can concede that humans' emotion recognition abilities correlate with certain personality traits, as psychology has noted, but this correlation might emerge more steadily with stimulus cues that are not in texts read out of context. An interesting course of action would be to only focus on a specific factor and diversify the types of events and linguistic productions (e.g., longer and more narrative reports of personal experiences) depending on that. This requires a radical re-thinking of the construction of an event corpus, with the generation phase leaving aside constraints about the variety of events (how different they should be) to gain more control over their semantics (what the events should be about); the validation phase might reduce the initial uncertainty about the annotators' personal characteristics by targeting specific groups of people, instead of merely asking for their characteristics after randomly assigning the texts to annotate, as we did. One could instruct all generators to elaborate on the same event, such as a fact that resonates with the history of a culture. This approach will yield multiple linguistic realizations of a single concept. Consequently, it will permit to test if validators with the same cultural roots (and only those validators) display similar emotion interpretations, and if the emotions patterns identified among respondents with certain personality traits correspond to ways in which the event has been recounted (e.g., they consistently map to recurring focal aspects in the narrations).

Second, alternative data analysis methods could identify more patterns than ours. We could have estimated the factors' systematic impact on the annotation performance via linear models. Multivariate regression models have been employed in many linguistic tasks (Beinborn et al., 2014; Papay et al., 2020; Dayanik et al., 2022) to find relationships between an observed (or dependent) variable, such as a system's performance, and one or more independent variables that covary with it. This is a promising approach to transfer on emotion and appraisal agreement. It allows predicting the performance of pairs of annotators through independent linguistic and extralinguistic variables considered all at once, without the need to create somewhat handcrafted subsets (see for instance our decision of whether the annotators belonged to the same *age* group) and to study them separately. That might be a more suitable way of describing each annotator's Θ. To be able and estimate the effect of these factors with sufficient statistical power, we would have needed more datapoints for each property, while our data was extremely sparse, as each validator of crowd-enVENT had a different combination of features and labeled no more than 5 texts.

Contextual and Linguistic Factors. In most cases, the performance of annotators in the same *emotion state* was more similar compared to that of subjects with an "emotion mismatch". We do not generalize this finding either, because the differences in measures of agreement between the two groups of people were modest. In fact, in the context of crowd-enVENT, they confirm that the task of putting oneself in the shoes of the writers promoted annotation quality – it smoothened out idiosyncratic affective associations guided by the coders' own affective condition.

Clearer results were obtained by tracing disagreement back to people's confidence and emotion intensity, two separate but interrelated sources of judgment. We investigated if the perceived emotion (class), a perceived feature of emotion (intensity), and a self-perception (confidence) are tied together and can help understand incongruent annotations. We found that the emotion recognition task elicits diverse decisions among the coders even in a simplified annotation setting, where raters distinguish *neutral* sentences from *emotion*-bearing ones. Both intensity and confidence accounted for inter-annotator divergence relative to those binary alternatives (RQ2). Adopting confidence as an acceptance threshold showed that higher scores lead to more uniform assessments of emotions; though not a surprising effect of confidence, this also applies to intensity. The finding is interesting in itself: some judgments which are seemingly unsolvable can be explained by certain perceived properties of the texts (emotion intensity) or self-perceived features (confidence). Moreover, (RQ3) the two variables are correlated, that is, people feel more certain about their emotion recognition performance on items with high intensity.

From these results we can draw some lessons. First, the correlation between confidence and intensity carries relevant implications for all those studies that focus on emotional strength. When asked to evaluate intensity, do people confound that with confidence? Even more, is there a causal relation between the two? As a best practice to put safeguards in their guidelines, experimenters may ask people to tease the two variables apart. This issue concerns modeling studies as well: Do regressors for emotion intensity predict such a dimension, or rather the confidence with which people judge emotions? We provided reasons to look into this question further.

Second, confidence turned out to be an important dimension of rating, because it can inform us when the annotators expect to disagree. When judgments diverge, annotators do not deem their intuition credible. Hence, our finding that confidence approximates disagreements means that people themselves predict their performance to differ from that of others. Emotion labels alone are not sufficient to unearth these details, which calls for annotation actions that address emotions as a multi-dimensional problem. For such a complex and at times fuzzy phenomenon, a well-supplied set of annotation variables is beneficial, as it exposes potential interactions between them.

Concretely, all this knowledge can come in support of annotation studies. Including confidence as a rating dimension gives additional information about the annotators' reliability. It can help experimenters refine the guidelines in a pre-testing phase: one might want to disentangle disagreements that are random, signaling a lack of annotators' reliability, from cases due to consistently different ways of perceiving (or reporting on) confidence. In the second case, disagreements can be normalized to an extent, by post-processing the annotation results. For instance, as people seem to agree on the class *emotion* only if they also agree on certain degrees of CONF, there might be levels of confidence (or intensity) that one filters in/out of the final annotation labels.

One pitfall of such a strategy is the loss of annotated perspectives, since only the items on which humans' intuition falls above a predefined threshold are accepted as emotional. While we observed agreement on those items, it is possible to adopt more nuanced evaluation approaches and integrate information about intensity or confidence into IAA measures. As an example, disagreements between two raters can be penalized more when the one choosing the *emotion* label does so by perceiving extreme confidence or intensity – even though we provided evidence that these cases are rare. In this vein, Bhowmick et al. (2010) used the variable of confidence in emotion annotation to quantify the coders' reliability through newly proposed fuzzy measures of agreement, where inconsistencies at higher confidence contribute more to the overall disagreement function. Future work could explore a similar direction.

What's Next for Emotion-based Annotation Studies? There are of course many other possible paths to develop annotation schemes while accounting for the multifaceted nature of emotions. If we are to ground "emotionally-intelligent" systems on human language use and understanding, at least three key improvements call for attention. One is more linguistic context – a future research direction where my discussion already landed in the first chapter. For the judges, texts larger than individual sentences can help infer why a target emotion expression is uttered; for the researchers, they might serve to explain what semantic information led the coders to link a certain label with that expression. Next is structured information that the annotators can produce. As an example, they can mark the portion of an utterance that convinced them to prefer a label over another. Lastly, multimodal information: this is an extralinguistic factor that describes, rather than the coders, the situation in which an utterance is communicated, providing a visual, auditory and proxemics-related anchors to understand annotations (e.g., by considering the emotion expression was produced by a certain character, with a certain pitch, in consequence to a certain action).

These three aspects have the potential to disambiguate emotions and possibly reduce disagreements. Other strategies can be devised, instead, to acknowledge that the emotion load of texts is not a rigid truth, as to "welcome" the plurality of (affective) interpretations that the texts elicit. Works could overcome a glaring simplification that my studies made: I only used self-reports by assuming the truthfulness of the coders' answers, their capability to assess and quantify their characteristics (e.g., How agreeable am I?), the validity of the used theories (e.g., whether personality traits are a proper way to characterize people), and the validity of the specific test used to measure them (see, e.g., Hogan et al. (1977) for an overview of favorable vs. skeptical arguments on the use of personality tests). While the annotations I collected might contain at least some useful facets of the coders, it is possible to complement them with much deeper information.

A way of doing so is to stretch the analysis of people's emotion recognition performances over time. Repeated measurements provide an overall more robust approach to investigate the effects of the Θ_i that contribute to an emotion recognition performance, since they reveal who chooses what label, when, what factors have changed from one measurement to another. This scenario allows studying the disagreements that emotions can spur not only between subjects but also within a coder herself. Plus, it permits to formalize Θ_i as a vector that includes more than self-reported ratings: the person producing the annotation would become a variable in its own right, and might explain some variance in the annotation that demographics, personality traits, and other manually-defined factors cannot capture.

This type of information would also promote large-scale analyses of agreement, for example by modeling the emotion decoding task: like human readers, a classifier can perform $\arg\max_{E \in \text{emotions}} p(\text{E} \mid \text{text}, \Theta_i)$. Hence, the assumption that pairs of annotators (j, k) with similar features agree more could be verified by looking at the match between the emotions predicted by the system when fed with the embedding of the text and the vectors of extralinguistic information regarding the annotators (Θ_j and Θ_k, one at a time).

These proposals exceed the sphere of text. But there is still ample room to study emotions while remaining in the verbal domain. Linguistic properties that supposedly pertain to emotions are what I am going to examine next.

Chapter 5

Is Emotion a "Style"?

The previous chapters described human-based studies. They addressed the questions of how well emotions are recognized by readers and what extralinguistic factors play a role in such a task. This chapter switches perspective. It uses computational tools to shed light on the linguistic level at which emotions emerge, as a first step toward understanding the cues that readers grasp to sense an affective import in text.

I focus on the question of whether emotions are recognized thanks to the style of an utterance. Experiments in a backtranslation setup show that (1) emotions are partially lost during translation; (2) this tendency can be reversed almost completely with a simple re-ranking approach informed by an emotion classifier, taking advantage of diversity in the n-best translation hypotheses; (3) the re-ranking approach can also be applied to change emotions, as a model for *emotion style transfer*. A qualitative analysis of the style-oriented paraphrases then reveals that emotions are toned down or amplified with recurring linguistic changes, but these changes signal successful transfer only if measured with automatic metrics, while presenting marginal emotion variations to the human eye.

1 Emotion as a Style to Transfer

Looking at style through theoretically agnostic lenses is routine in the style transfer literature. The hierarchy in Chapter 2 (Page 53) visualized the linguistic features treated as styles in the field, and while it

Highlights

Emotions tend to diminish during (back)translation.

If style is a dimension of language orthogonal to meaning, emotions are not stylistic attributes.

By obtaining many back-translation hypotheses, a paraphrase with a target emotion (according to classifiers) can be found.

Different emotion expressions might need different transfer treatment.

underlined their commonalities (the high-level communicative characteristics along which the hierarchy ramifies), it also presented how diverse "styles" are. Some, like politeness, are tailored to one's interlocutor and have clear linguistic markers, others are influenced by the environment (e.g., the style of the time), and yet others depend on the characteristics of the person producing an utterance (e.g., age). The publications' authors refrain from providing a conclusive definition of the phenomenon they look at. They rather decide on an ad-hoc function that separates language into sets of mutually exclusive attributes. For instance, verbal expressions can be either formal or informal, and in that case, style would be the partition function of formality.

An operational framework on which the works converge, however, can be identified. Style, which has to be altered, is a dimension of language orthogonal to meaning, which instead has to be kept invariant. Meaning is a bit of information that could be decoded from different surface realizations, and style is the set of linguistic choices that encode it. Hence, to summarize the typical approach to the style–content binomial, one could say that style is all that can be transferred by style transfer algorithms. I follow through with this operational paradigm, but putting the accent on the style rather than the transfer. That is, instead of assuming that since X is a style then it can be transferred, I concede that if X can be transferred then it is a style. This will allow me to verify the assumption that emotions can be considered as the form we give to meanings.

Past style transfer research has mostly focused on binary attributes,

Anger:	I'll play a song over my dead body.
Sadness:	I wish I could play a song, sorry.
Anger:	Sit down.
No Emotion:	Please have a seat.

Figure 5.1: Emotion style transfer: Anger to Sadness and Anger to No Emotion.

like positive vs. negative polarity, formal vs. informal attitude. Very few works have moved to the variable of emotion, which involves a multitude of attributes (e.g., the six emotions of Ekman) interacting with one another in different ways. Intuitively, a sentence written with an angry tone is more difficult to reformulate into a joyful one than a fearful one.

An example of emotion transfer is in Figure 5.1. To understand its inherent challenges, Helbig et al. (2020) adopted a strategy for explicit style-to-content disentanglement. They designed a pipeline that sequentially determines the portion of text to modify, performs the change, and filters out suboptimal paraphrases by scoring their content preservation, presence of the target emotion, and fluency. In contrast to a black-box neural encoder/decoder approach, this setup enabled them to investigate the point in the lexical substitution process where changing emotional attributes becomes a gruelling chore. Their qualitative evaluation further showed that the simultaneous adjustments of content and emotion are conflicting objectives, especially for implicit emotion expressions of cognitive appraisals or descriptions of bodily reactions.

This preliminary study gives reason to carry the task over in other directions. Targeting individual words for the edits turned out little productive, and satisfying the three desiderata arduous. A scheme that facilitates their achievement, without requiring to handcraft three objective functions, would be ideal. I obtain that with a machine translation-based approach. Further, as a follow-up to the finding that the transfer is not equally straightforward with all expressions, it is reasonable to investigate various linguistic realizations of emotions, to gain a better grasp of the cases in which they can be transferred. I collect this insight by using multiple datasets for emotion analysis, built in a range of textual domains.

1.1 The Link between Style Transfer and Machine Translation

What differentiates style transfer and machine translation (MT) is that the former pursues a mapping of attributes, while the latter operates a language-to-language transformation. Still, there is a marked resemblance between the two tasks: both aim at the objectives of content preservation and text fluency. Translation can actually be thought of as a kind of transfer that captures the crux of the task (rewriting the same content in a new attribute, i.e., a new language), despite due differences between the style “language” and others – e.g., it has many more attributes (English, German, Spanish, etc.); the transfer strength is not to be measured; people can vary the language in which they communicate at will, but hardly in the same way as other intentional leaves in the hierarchy I described.

In analogy to style transfer systems, MT offers a viable solution for the automatic generation of paraphrases, namely, backtranslation (Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Bannard and Callison-Burch, 2005; Mallinson et al., 2017; Wieting and Gimpel, 2018). Mapping from a source language S to a target language T and back (i.e., $S \to T \to S$) creates texts that are fluent in S, maintain the semantics of the input, and likely have a different surface form. Indeed, (human and automatic) translation often induces structural changes in the properties that characterized the input text. For example, multi-word expressions can be collapsed into single-word terms (Zaninello and Birch, 2020), and verbs can be passivized (He et al., 2015).

Put differently, since MT systems are optimized for both output fluency and adequacy between input and output, using them in a backtranslation framework produces paraphrases that satisfy the transfer desiderata of naturalness and content preservation.

1.2 Preserving vs. Transferring Emotions

The quest for a MT-based method to transfer emotions first requires understanding if emotions undergo any changes during translation, as a preliminary step to applying style transfer-oriented interventions on the MT systems.

Nowadays, fluency and adequacy are met more often than not, and in those respects, the quality of MT in some areas follows close behind that of people (Barrault et al., 2019), so much so that translation models are deployed in downstream NLP tasks such as sentence simplification (Xu et al., 2016), error correction (Yuan and Briscoe, 2016), and cross-lingual resource creation (Barnes and Klinger, 2019). They also enable human-to-human communication across languages, e.g., in chat systems, customer support, or social media. On the other hand, the way automatic translators handle subtle and extra-propositional information is far from perfect. They proved to obfuscate many stylistic properties, in particular some socio-demographic characteristics of the texts' authors, like gender and personality traits (Mirkin et al., 2015; Rabinovich et al., 2017).

As far as emotions go, the extent to which MT maintains them is unclear. Current translation systems do not comprise any special component to process affective phenomena. Nor do they ensure, next to content adequacy, an "emotion adequacy". They *should* keep emotions unaltered: if a source text expresses an affective import, the MT user would expect a translation to convey the same, because emotions are an aspect that creates common ground in machine-mediated communication (Yamashita et al., 2009). Yet, translations are subjected to a series of linguistic constraints specific to the languages under consideration, like the absence of terms for certain states (e.g., *Sehnsucht* is German for a longing for some absent thing) or colexification phenomena (i.e., naming related emotions with the same word, like *grief* and *regret* in Persian) which vary from language to language (Jackson et al., 2019). Moreover, aesthetic considerations often call for making texts readable or pleasant more than literally faithful to the input. All these factors can cause not only translation errors (for human and machines alike) but also stylistic choices that subvert, for example, the sentiment of words (Petrova and Rodionova, 2016; Saadany and Orasan, 2020). It is possible that MT amends emotions as well – even more remarkably in backtranslation, where the linguistic transformation happens twice.

Preserving emotions, affect and sentiment in text is an issue for cross-linguistic studies (Wierzbicka, 2013; Wassmann, 2017; Hubscher-Davidson, 2017) which demands an assessment of the quality of data obtained by translation (Banea et al., 2008; Chen and Skiena, 2014; Buechel et al., 2020, i.a.). By validating resources for Romanian derived from English ones, Mihalcea et al. (2007) noticed that human translation can obscure the subjectivity of a lexicon. A similar observation was drawn for polarity by Balahur and Turchi (2012) with statistical MT, and by Salameh et al. (2015) and Mohammad et al. (2016b) who found that translation can corrupt textual sentiment, flattening positive and negative aspects down to neutrality. Hence, translated polarity, subjectivity, valence, dominance, and arousal have attracted works in the past, but the case of emotions in MT has gone rather undiscovered. Kajava et al. (2020) compared parallel movie subtitles in English, Finnish, Italian, and French and concluded that emotion preservation depends on the language pair; however, the origin of those subtitles (e.g., automatic systems, professional human translators, amateurs) was uncertain.

Some studies tried to incentivize the consistency of sentiment between input and output. Lohar et al. (2017) built separate translation models for data coming from different sentiment categories. Si et al. (2019) directly incorporated sentiment in their neural MT system, implementing a Seq2Seq English-to-Chinese translation model that keeps not only the semantics but also the sentiment of the input, both by including the sentiment label in the source sentences, and by learning the negative/positive meanings of ambiguous words as separate embeddings.

The idea that stylistic attributes are less distinguishable after translation had a different effect on style transfer, where it gave momentum to the attempt of promoting (rather than concealing) the shift of attributes. Prabhumoye et al. (2018) used MT as a paraphrasing and implicit disentanglement solution to transfer the styles of gender, political slant and sentiment. A sentence in the source language was translated into a pivot language; encoding the resulting text in a backtranslation step then served to produce a style-devoid representation for it; the final generation step consisted in decoding the style-less representation towards a specific attribute, with the help of one decoder per attribute.

Goal of the Study. This chapter has a twofold aim. One is to come to grips with what happens to emotions during translation. The other is to take advantage of the re-wordings that characterize translation outputs, to make these display a desired emotion attribute, which could be the same or different from that contained in the input.

First, I set the style transfer goal aside, and using state-of-the-art off-the-shelf systems, I investigate how well emotions are preserved in MT (RQ1). Fixing the knowledge gap on this issue is important theoretically (to inform cross-lingual studies that use translation as part of their experimental setup) and practically (both to improve the usefulness of MT and to better understand the possibilities of MT for style transfer).

Research Question	Sentence
	He was furious at the apparent disregard for rules.
RQ1: Does MT dilute emotions?	He was worried at the apparent disregard for rules.
RQ2: Can we recover the input emotion?	He was quite enraged at the inattention to rules.
RQ3: Can we change the emotion?	He was unhappy that the rules were ignored.

Table 5.1: Illustration of three emotion-related research questions about MT, with examples for the associated tasks to be solved. The text in bold is the input.

As I find that emotion nuances tend to vanish in translation like other extra-propositional properties, I proceed to counteract their fading (RQ2). I establish an emotion-based translation candidate re-ranking strategy that is applied as post-processing to an MT system's n-best output. Lastly, I address the core focus of this chapter and consider how re-ranking can be used for style transfer (RQ3). The potential finding that such a transfer is doable would permit me to conclude that emotions emerge in language in the guise of a style.

Table 5.1 showcases example outputs from the experiments of each research question. Given an input, backtranslation produces a text with an emotionally-connotated adjective ("worried") that moves the sentence away from the original affective load (RQ1); that is then restored with our re-ranking procedure (RQ2) and turned into a different emotion with the same approach, for style transfer (RQ3).

2 A Method for Emotion Preservation and Transfer in Neural Machine Translation

Achieving the transfer through translation methods is convenient for various reasons. It only requires controlling the output attribute (cf. Prabhumoye et al., 2018), while the other two criteria of fluency and similarity to the input content are optimized by the MT models. Translation further diverts attention from the explicit disentanglement of content and style that was unsuccessful in Helbig et al. (2020), since the

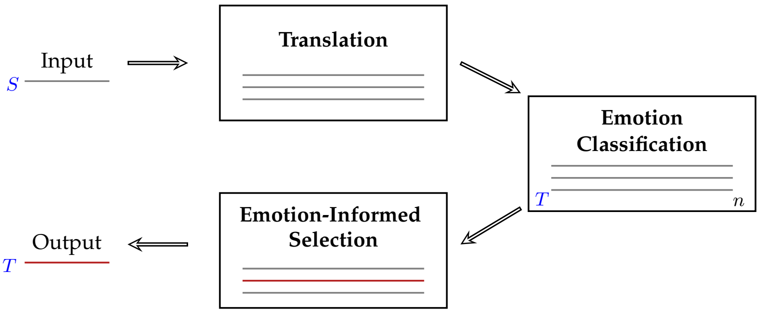


Figure 5.2: Overview of emotion preservation and transfer method. "S/T": source/target language; "n": number of translation hypotheses. The red line is the best candidate returned as output.

transformations that MT makes on the textual surface are not necessarily word-by-word substitutions (Barzilay and Lee, 2003). However, similar to Helbig et al. (2020), we aim at creating a method transparent to investigation, without the need for different decoders dedicated to each attribute, neither for a method that produces the style–content separation on the texts' latent representations, nor for embeddings of the attributes in question to promote the desired decoding (as in, e.g., Smith et al., 2019).

To that end, we conceptualize both emotion preservation and emotion style transfer in neural MT (NMT) as a post-processing re-ranking step. The re-ranking can be defined on the basis of a standard emotion classifier with probabilistic output, to select a candidate whose emotional connotation is as close as possible to the input (for preservation) or is arbitrarily decided (for the transfer).

As shown in Figure 5.2, our pipeline involves three components: a translation model, an emotion classifier, and a candidate selection procedure. Starting from an input in the source language S, we generate the n-best translation candidates in a target language T with an NMT system, which is presumably agnostic to emotion-specific considerations. Then, we re-rank these candidates based on probabilities produced by an emotion classifier, and select the best hypothesis given those emotion-level considerations. Hence, the crucial variable is the *diversity* of the n-best list: the more diverse, the better the emotion classifier can promote hypotheses that express particular emotions, even if they are not optimal from the point of view of the overall scoring function of the NMT system. Below, we introduce the three pipeline components.

Translation Model. We require a translation model that returns a list of n-best translation candidates, which is the case for essentially every statistical or neural MT system. We use FAIRSEQ (Ott et al., 2019), an open-source sequence-to-sequence modeling toolkit applicable to various tasks, MT included. It shows state-of-the-art performance and it was developed with the goal to replicate different model architectures.

Importantly, FAIRSEQ supports different search algorithms, like beam search and top-k sampling, which differ in their ability to encourage diversity in the output. Beam search searches the space of hypotheses left-to-right, retaining at each time step a number of top-scoring candidates that equals the width of the beam, and expanding on those. Sequences decoded with beam search differ on minimal portions (Gimpel et al., 2013), while they are more varied when generated with sampling strategies. Top-k sampling, for instance, does not aim at maximizing the likelihood of text. Instead, it randomly samples words step-wise and outputs from the top-k most probable ones (Fan et al., 2018).

Emotion Classification Model. To estimate the probability distribution over emotions for a given text, we use a BiLSTM with a self-attention mechanism. This model architecture does not correspond to the one used in Chapter 4, but it appears equally promising, as it has been shown to perform close to state-of-the-art in emotion analysis (Baziotis et al., 2018). We treat the output of the classifier as a scoring function emo$(t, e) = p(e \mid t)$, i.e., the conditional probability of an emotion given a text t, and we assume that it is comparable across languages (see Section 3 for a discussion of this assumption).

The discrete emotion labels for which the probability distribution is produced represent the attributes for style transfer in RQ3.

Translation Candidate Selection. Once the n translation candidates (called $\mathrm{hypotheses}$ in Equations 5.1 and 5.2 below) are scored by the emotion classifier, we re-rank them based on their probability for specific emotions, and select a top candidate based on our research question. This can be done with the setup described above: it suffices to adjust the selection strategy for each research question separately.

In RQ1, where we only consider a single translation hypothesis, the output selected by emotion selection is trivially the one coming out of the translation — it is picked based on properties of a standard translation procedure.

For RQ2, we preserve the dominant emotion of the input by selecting an output such that

$$\texttt{output} = \underset{c \in \text{hypotheses}(\texttt{input})}{\arg\min} |\text{emo}(c, \hat{e}) - \text{emo}(\texttt{input}, \hat{e})| \tag{5.1}$$

where $\hat{e} = \arg\max_{e \in \text{Emotions}} \text{emo}(\texttt{input}, e)$. Measuring the absolute difference in emotion load for two texts is similar to Luo et al. (2019), who analyzed the change in sentiment intensity with mean absolute errors.

Finally, in RQ3, where we aim at maximizing an arbitrarily chosen emotion e', we define

$$\texttt{output} = \underset{c \in \text{hypotheses}(\texttt{input})}{\arg\max} \text{emo}(c, e') \,. \tag{5.2}$$

Our method does not condition the MT system towards a specific emotion. Instead, we evaluate the extent to which the n-best lists of a state-of-the-art MT system contain sufficient variation in their candidates as to manipulate the emotional load of a translation – either by optimizing the preservation of the input emotion (RQ2) or by changing the emotion connotation (RQ3).

3 Experimental Setup

The most opportune setup to carry out the re-ranking for emotion preservation (RQ1) would be bilingual, to analyze the translation of some source language text into a target language. For instance, one could compare the distribution of emotion probabilities for a translation against the corresponding distribution for the source text. However, a meaningful cross-lingual comparison of emotion probabilities is methodologically challenging: this would require either manual annotation or highly comparable emotion classifiers for the source and target languages. Manual annotations are costly, and emotion annotation can be tricky in terms of inter-subjective replicability (cf. Chapter 3 and Chapter 4). Neither are we aware of emotion classifiers with evidently similar behaviors across languages.

Backtranslation enables us to avoid the issue of cross-lingual comparability (Mallinson et al., 2017): instead of analyzing the translations

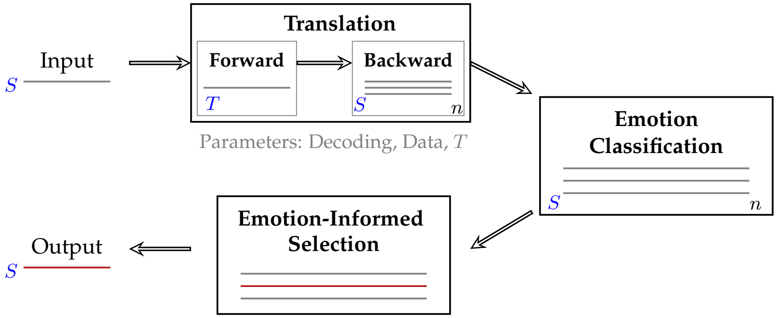


Figure 5.3: Instantiation of our methodology with backtranslation. "Parameters" under the box "Translation" are those we vary in the experiments.

obtained from a system performing S(ource)$\rightarrow T$(arget)[1], we consider the output of the mapping $S \rightarrow T \rightarrow S$, that we can examine with only one emotion classifier for the source language. Hence, we instantiate a backtranslation version of the method described above. This setup is shown in Figure 5.3. Given an input in S and a target emotion, we generate the best translation in T. We then translate it back into multiple hypotheses, obtaining a set of n paraphrases for the input.

Regarding the first research question, our solution assumes that experimenting with backtranslation gives a realistic picture of what would happen in a $S \rightarrow T$ setting. This is a simplifying assumption, which does not measure the loss of emotion in one direction, nor accounts for *where* the change in emotion occurs, but adding a translation step seems a reasonable compromise in the absence of comparable emotion classifiers for different languages. First, while the *absolute* magnitude of the problem might be overestimated, with our monolingual classifiers we can still look at the *comparative* magnitude of emotion loss across different MT settings, which we do in the sections below. Results indicating that we can *improve* emotion preservation in backtranslation (RQ2) would be stronger than such results obtained on a single translation step. Second, backtranslation is necessary to create paraphrases of an input for style transfer; and if it were to make extreme changes on the surface of texts (which propagate from the first to the second translation step), that would be disadvantageous to study emotion preservation for RQ1, but a favorable condition to find emotional variations for style transfer in RQ3.

[1] T is target langage, different from target emotion attribute.

Following the considerations above, we do not run a single experiment, but carry out a series of comparisons, varying the different parameters of our method.

NMT Model: Varying Target Language and Sampling Method. We use FAIRSEQ with English–German and English–Russian models[2] (Ng et al., 2019). These sentence-level models are based on transformers (Vaswani et al., 2017). They are pretrained on bitext and backtranslated news data, fine-tuned on in-domain data and used for decoding with a noisy channel approach to re-rank the n-best hypotheses. We use these models both with beam search and top-k sampling (cf. Section 2).

Data Sets: Varying Emotion Realization. As highlighted in Chapter 2, emotion realizations differ across domains and genres. To gain a representative picture and investigate the effect of translation on different emotion realizations, we compare four English corpora: ISEAR (Scherer and Wallbott, 1994), whose $\approx$7k event descriptions are labeled with the emotion that the events induced in the experiencers (*anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*); TEC (Mohammad, 2012), where $\approx$ 21k tweets are associated to the six fundamental Ekman's emotions (Ekman, 1992); the corpora by Aman and Szpakowicz (2007) and by Alm et al. (2005), repertoires of $\approx$5k and $\approx$15k sentences from a number of Blogs and (fairy-)Tales, respectively, using Ekman and *no emotion* (*noemotion*). These corpora differ in labels (see Figure 5.4 vs. 5.5), topics, registers and communicative purposes: TEC collects short, spontaneous expressions, ISEAR provides statements produced in lab.

Emotion Classifier. Due to these differences in linguistic realization among corpora, emotion classifiers generalize badly (Bostan and Klinger, 2018). To avoid this problem, we re-train our emotion classifier (cf. Section 2) for each dataset. We train the model on 70% of the instances (cf. Section 3), validating it on the 10% and using the remaining 20% to evaluate our emotion preservation method. We use 300-dimensional GloVe embeddings (Pennington et al., 2014); for regularization, we use Gaussian noise, a dropout rate of 0.1, and early stopping. Table 5.2 shows that performance on the various corpora is comparable to previous work on the same setup (Bostan and Klinger, 2018).

[2] `https://github.com/pytorch/fairseq/tree/master/examples/wmt19`.

	Micro F_1			
Emotion	ISEAR	Blogs	Tales	TEC
Anger	.51	.55	.39	.37
Disgust	.58	.64	.12	.26
Fear	.70	.56	.33	.55
Guilt	.55	—	—	—
Joy	.72	.69	.45	.69
No-Emotion	—	.88	.79	—
Sadness	.61	.49	.37	.45
Surprise	—	.41	.27	.49
Shame	.46	—	—	—

Table 5.2: Classification results ("—": the emotion is not a label in the respective corpus).

Evaluation. For evaluation, we re-use the emotion scores employed in candidate ranking. Our basic measure is again based on probability differences with regard to a specific emotion e in a set S of input–output pairs:

$$\Delta(S, e) = \frac{1}{|S|} \sum_{(s_1,s_2)\in S} \text{emo}(s_2, e) - \text{emo}(s_1, e). \tag{5.3}$$

For RQ1, S is the set of inputs (s_1) and their 1-best backtranslations (s_2). For RQ2, S is the set of inputs and their backtranslations as selected by Eq. (5.1) for each input emotion. In RQ3, S is the set of inputs and their backtranslations as selected by Eq. (5.2) for each target emotion.

This measure ranges in [-1,1]. Its advantage is that it can be applied to all research questions to observe how an emotion e has changed. Δ = 0 means that the score that an emotion is assigned in the distribution of probabilities produced by the classifier is exactly the same in the output and the input. A positive score signals that the output has more of that emotion than the input. Vice versa for negative scores. Δ = 1 and Δ = -1 indicate that the input and output maximally differ with respect to the emotion in question.

We acknowledge that using the emotion classifier both for ranking and evaluation introduces a potential circularity. To avoid this problem, the reliability of the classifier is crucial. We therefore carry out a qualitative inspection of examples (Section 5) to compare the classifier output against our judgments.

4 Results

We now answer the three research questions, separately.

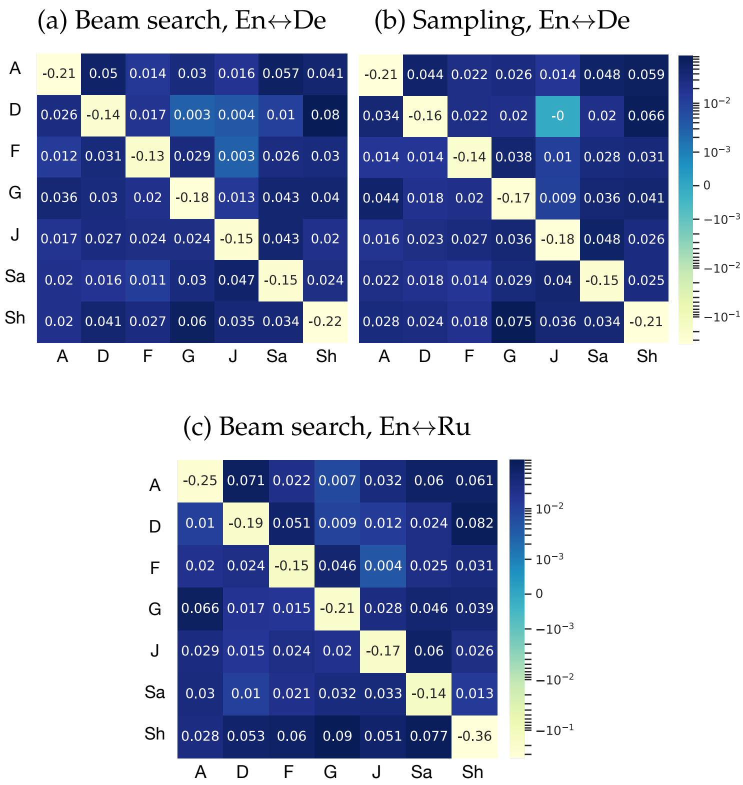
4.1 RQ1: Does Translation Preserve the Emotion Connotation of Texts?

Let us first question if the off-the-shelf system FAIRSEQ reduces emotion connotations. This analysis is based on the $n = 1$ best output from the translation system, which we compare to the original input.

Figure 5.4 and Figure 5.5 show the Δ values between the input and output emotion probabilities. Each cell in the heatmaps contains the average difference between the group of texts that are associated with the emotion on the row (as determined by our emotion classifier) and their backtranslations. For instance, the first row in Figure 5.4 informs us about the extent to which the emotions on the columns change when input texts expressing predominantly *anger* are backtranslated: the probability is reduced by an average of 21% for *anger*, while it increases a bit for all the other emotions. Our expectation that backtranslations have a lower emotional score for the emotion characterizing the input should reflect on the diagonal, which reports the Δ values between the emotion identified by the classifier in an input text and the same emotion as measured in its backtranslation.

In order to establish what patterns have general validity, we vary three parameters (cf. Section 3 for details): the *data set* (ISEAR, TEC, Tales, Blogs), to measure the influence of domain and annotation procedure; the *language* in the forward translation step (from English to German and from English to Russian); and the *decoding strategy*, comparing beam search, which is more conservative, to sampling, which generates more diverse results.

Varying Decoding Method and Target Language. We analyze decoding method and target language on ISEAR. Figure 5.4 reports the results obtained when using beam search (a) against sampling (b), and German (a) against Russian (c). There is no significant difference between German and Russian (p = .23, Mann Whitney U test), nor between decoding methods (p = .76). We conclude that the ability of translations to preserve emotion is unrelated to the target language and the generation strategies we employed.



(a) Beam search, En↔De

	A	D	F	G	J	Sa	Sh
A	-0.21	0.05	0.014	0.03	0.016	0.057	0.041
D	0.026	-0.14	0.017	0.003	0.004	0.01	0.08
F	0.012	0.031	-0.13	0.029	0.003	0.026	0.03
G	0.036	0.03	0.02	-0.18	0.013	0.043	0.04
J	0.017	0.027	0.024	0.024	-0.15	0.043	0.02
Sa	0.02	0.016	0.011	0.03	0.047	-0.15	0.024
Sh	0.02	0.041	0.027	0.06	0.035	0.034	-0.22

(b) Sampling, En↔De

	A	D	F	G	J	Sa	Sh
A	-0.21	0.044	0.022	0.026	0.014	0.048	0.059
D	0.034	-0.16	0.022	0.02	-0	0.02	0.066
F	0.014	0.014	-0.14	0.038	0.01	0.028	0.031
G	0.044	0.018	0.02	-0.17	0.009	0.036	0.041
J	0.016	0.023	0.027	0.036	-0.18	0.048	0.026
Sa	0.022	0.018	0.014	0.029	0.04	-0.15	0.025
Sh	0.028	0.024	0.018	0.075	0.036	0.034	-0.21

(c) Beam search, En↔Ru

	A	D	F	G	J	Sa	Sh
A	-0.25	0.071	0.022	0.007	0.032	0.06	0.061
D	0.01	-0.19	0.051	0.009	0.012	0.024	0.082
F	0.02	0.024	-0.15	0.046	0.004	0.025	0.031
G	0.066	0.017	0.015	-0.21	0.028	0.046	0.039
J	0.029	0.015	0.024	0.02	-0.17	0.06	0.026
Sa	0.03	0.01	0.021	0.032	0.033	-0.14	0.013
Sh	0.028	0.053	0.06	0.09	0.051	0.077	-0.36

Figure 5.4: RQ1: Emotion loss (Δ) on ISEAR, found with different parameter configurations. Rows are input emotions, columns are the output emotions (A: anger, D: disgust, F: fear, G: guilt, J: joy, Sa: sadness, Sh: shame). Each row shows the average Δ in per-class emotion output.

The values on the diagonals indicate a general loss of the dominant emotion in the input. They have the lowest magnitude and are negative. The backtranslations of inputs expressing *anger* and *shame* are those with the greatest loss in those same emotions ($-.21$ and $-.22$, respectively), followed by *guilt* ($-.18$), *joy* and *sadness* ($-.15$), *disgust* and, lastly, *fear* ($-.14$ and $-.13$). Instead, off-diagonal cells are positive, with the exception of the degree of *joy* in items originally containing *disgust*, which falls sligthly below 0 when the decoding is sampling. In the three cases, the highest increases are recorded for the instances originally labeled as *disgust*, which have an increase in their *shame*

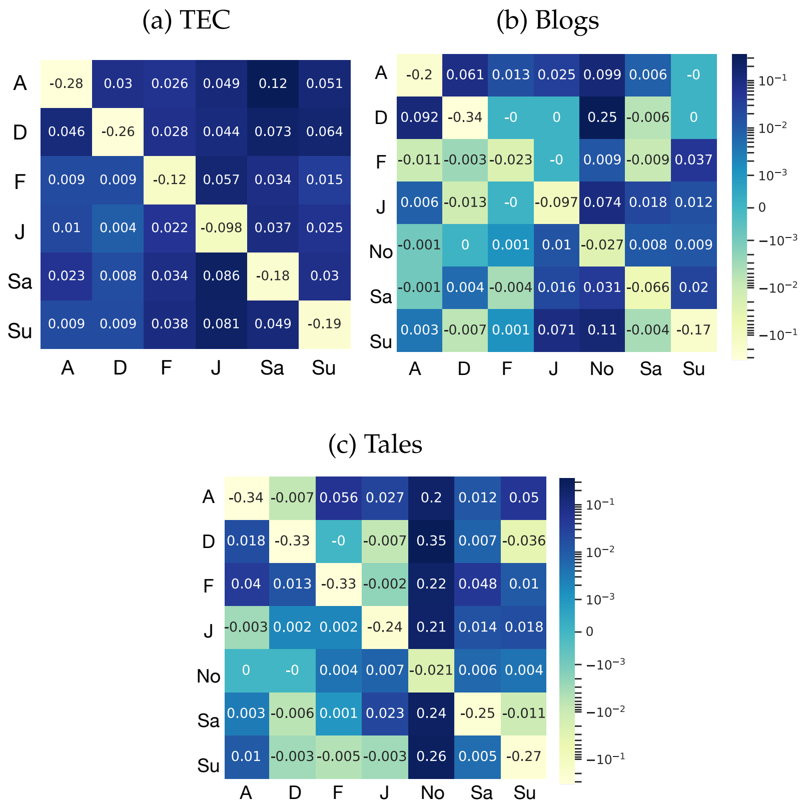


Figure 5.5: RQ1: Emotion loss on TEC, Blogs and Tales, using beam search for decoding and En↔De as language pairs. Rows are input emotions, columns are the output emotions (No: no emotion, Su: surprise).

scores, and for the *shame* examples, whose amount of *guilt* is scaled up.

Overall, these numbers confirm our hypothesis: (back)translations express the original emotion to a lower extent than the input texts. The decrease of that emotion is balanced out by an increase of the others.

Varying Corpora. Given the non-significant difference between the parameters we tested, we continue our experiments by setting the decoding method and language pair to beam search and En↔De. We investigate if we can generalize our observations to datasets other than ISEAR.

Results are reported in Figure 5.5. They suggest that the loss of original emotions in the diagonal is a persistent trend across corpora, together with the fact that the predominant input emotions are toned down more than any other. We observe that the emotion change on TEC is the most similar to ISEAR, despite the difference in their labels. Further, it is interesting to compare the amount of *anger* gained by the translations of texts classified as *disgust* in Figure 5.4 (a) vs. Figure 5.5 (b). This could be an effect of the presence of the label *noemotion*, which does not exist in ISEAR. It is also noticeable that translations of Blogs and Tales tend to increase in neutrality more than in other emotions. Exceptions are translations that were already classified as containing *noemotion*, and which lose their neutral status (see cell *noemotion-noemotion* in the diagonals of Figure 5.5 (b) and (c)).

4.2 RQ2: Can an Emotion-Informed Translation Selection Recover Input Emotions?

We can now evaluate the emotion-informed post-processing. For an input, we obtain its forward translation[3] and $n = 50$ backtranslations; among them, we pick the one minimizing the Δ with the input emotion, following Eq. (5.1).

Figure 5.6 (a) reports the results on ISEAR with beam search. Like before, the emotions on the rows are those expressed by the input text. Columns are the emotions for which the delta is computed between output and input. For instance, the cell A-D shows the average Δ between the *disgust* score of the texts classified as *anger*, and the *disgust* score in their backtranslations. What interests us is again the diagonal, showing the average differences between the original emotion and that emotion as estimated in the output. Once more these values are negative, indicating that at least for some texts, the translation with the closest emotion to the original one still has less of that emotion. Since this second research question required minimizing the deltas, values close to 0 indicate success. Most are actually close; the cells that depart from 0 the most are A-A, Sh-Sh and G-G, with $\Delta = -.042, -.042$ and $-.022$. In comparison to Figure 5.4 (a), we see that we can recover emotions. The loss of *anger* (A-A) is 5 times smaller than it was when exploiting the 1-best backtranslation; likewise, *sadness* (Sa-Sa) is preserved $\approx$21

[3] Leveraging more translations in T yielded results comparable to our current setting. However, focusing on n = 1 is advantageous for us, because it limits the artefacts introduced by *back*translation, and better approximates a more realistic setting in which the mapping between source and target occurs in a single step.

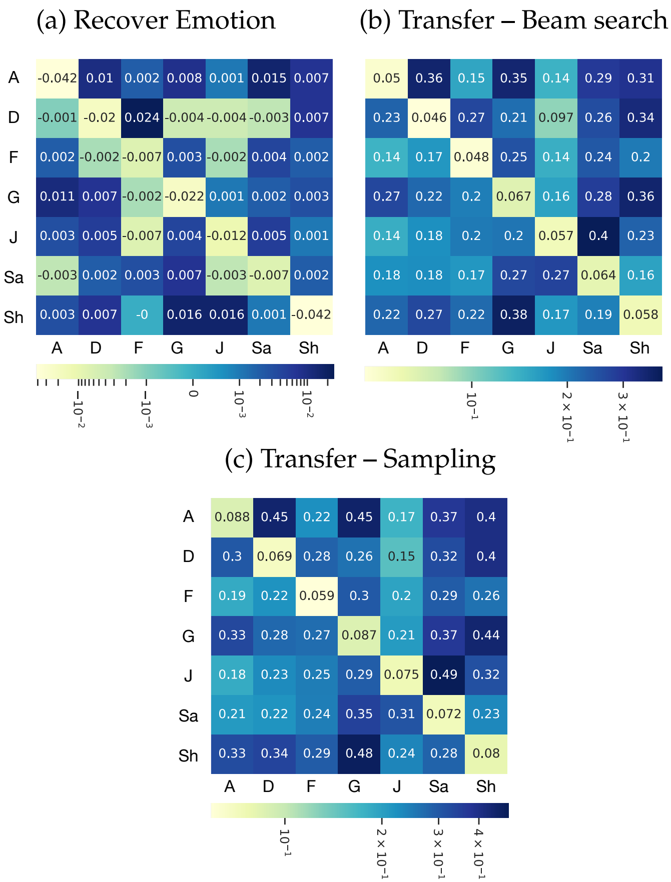


Figure 5.6: RQ2 and RQ3. RQ2: Heatmap (a) Recover Emotion reports the Δs for the second experiment. RQ3: Heatmaps (b) and (c) report the Δs for the third. In both cases, the dataset is ISEAR, input emotions are on the rows, columns are target emotions. See Figure 5.4 for emotion abbreviations.

times more. These numbers suggest that the behavior of NMT tools can be improved with the n-best lists produced by the systems themselves, which provide enough information to preserve emotions.

As a sanity check, we investigate if descending the n-best list in the search of an emotionally adequate translation had an impact on translation adequacy. We compare the average BLEU-4 score for the top outputs returned by the system (i.e., those analysed in RQ1) and our emotion-preserving backtranslations. Translation quality remains stable: in the first case we obtain .49 BLEU, in the latter we find a BLEU of .51. This indicates that it is possible to find an emotion-preserving variation further down the space of candidate outputs (at least to a certain point) without sacrificing the system's performance.

4.3 RQ3: Can Overgeneration Transfer Target Emotions?

Having shown that MT prefers to output sentences with a toned-down emotion, and that it is possible to subselect instances with a similar emotional connotation as the input, we now use the backtranslation pipeline as a paraphrasing tool for emotion style transfer, investigating if the diversity in MT output is functional for the task.

Given an input text t and a target emotion e, our goal is to produce a variation t' which is semantically similar to t, is fluent, and has the attribute e. The adopted MT systems are already trained to maximize the fluency of their output and their faithfulness to the input. Therefore, we assume that it is sufficient to pay attention to the presence of the target emotion (see Eq. (5.1)).

Forward and backward translation steps alike are carried out via beam search or top-k sampling, with k = 10, both producing n = 50 paraphrases. Since this experiment tries to promote stylistic diversity, n-best lists could be leveraged also in the target language. We experimented with multiple target translations as well, and we obtained results similar to those reported here, in which we employ only one forward hypothesis.

Varying Decoding. Figure 5.6 shows the results on ISEAR with beam search (b) and sampling (c). Each cell reports the average Δ of all instances for a pair of input (rows) and target (columns) emotions. It quantifies the strength of the transfer, or how much more of a target emotion is present in the selected backtranslation. For example, the first row in (b) considers the backtranslations of texts expressing *anger*. Those on which *anger* itself was transferred (i.e., those selected as having the highest degree of *anger*) express that emotion $.05$ points more than their original counterparts; *disgust* is $.36$ points higher than before in the backtranslations for which *disgust* was the target emotion.

As expected, the diagonal has the lowest numbers in both matrices, since it corresponds to target emotions that were already in the inputs.

Overall, there is quite a substantial emotion increase, indicating that our method can be used for the transfer of fine-grained affective attributes. The highest Δs are mainly among pairs of negative emotions. We also notice that it is easier to transfer *joy* onto negative emotions than the other way around (see column *joy*, which has some of the smallest off-diagonal values). This suggests there are interdependencies between the source text emotion and the desired target emotion.

Indeed, the transfer strength depends on input and target emotions in both the beam search and sampling cases. Successful transfers take place for sentences originally labeled as *joy* that are re-phrased as *sadness*. Given *shame*, *guilt* can be increased to a considerable extent, as can *shame* given *guilt*, which is a striking symmetry because these two emotions are attributed to properties of who experiences them (Tracy and Robins, 2006). Other than these similarities, however, we find a significant difference between the two matrices (p = $1.11 \cdot 10^{-09}$, Mann Whitney U test). The higher numbers in (c) corroborate the idea that sampling induces diversity in the n-best outputs more efficiently.

Emotion diversity in the translations can also be variously achieved by considering hypothesis spaces of different sizes. In heatmap (c) the diagonal mean is $.05$ and the off-diagonal 0.2; with n = 20 paraphrases, the diagonal decreases to a mean of $.04$ and the off-diagonal to $.18$; with n = 100, the diagonal and off-diagonal means are respectively $.09$ and $.39$. In sum, a higher n enables a stronger transfer.

5 Qualitative Analysis and Discussion

To gain further insight into our procedure, we show some instances from ISEAR that we found challenging for our models. Table 5.3 reports them across the three experiments in the beam search scenario (letters in bold correspond to inputs; additional examples are in Appendix, Table B.12). Their backtranslations lost the original connotation, so much so that the classifier assigns them to a different emotion class (this happens for 387 inputs in setup (a), see Figure 5.4).

Emotion variations (both loss and alteration) often involve a relatively small number of recurring linguistic transformations, like changes in modality (c. and f.) or in the intensity of the adjectives (b. and d.). The fact that *disgust* leaves room for *shame* (c.) appears coherent with the theory that the latter is related to self-responsibilities (Tracy and Robins, 2006): as opposed to the output of the transfer, the

RQ	Emo.	Sentence
	A	**a.** *When I have to take exams I am very excited and have not much time for the housekeeping. Then my friend has to do everything*
1	G	When I have to take exams, I am very excited and do not have much time for the budget. Then my boyfriend has to do everything.
2	A	When I have to take exams, I am very excited and have little time for housekeeping. Then my girlfriend has to do everything she can.
3	J	When I have exams, I'm very excited and I don't have much time for the household. Then my boyfriend has to take care of everything.
	A	**b.** *When a friend told me a story and I stayed dumb because I had no story to tell.*
1	F	When a friend told me a story and I remained silent because I had no story to tell.
2	A	If a friend told me a story and I was mute because I had no story to tell.
3	G	When a friend told me a story, I stayed silent because I had nothing to tell.
	D	**c.** *On New Years eve I drank too much alcohol, so much that I had to vomit in the presence of other people.*
1	Sh	On New Year's Eve I drank so much alcohol that I vomited in the presence of other people.
2	D	On New Year's Eve I drank so much alcohol that I had to vomit in the presence of other people.
3	G	On New Year's Eve, I drank too much alcohol, so much that I threw up in the presence of other people.
	D	**d.** *I feel disgusted with the bootlickers, with helpless people.*
1	Sa	I loathe the bootleggers, the helpless people.
2	D	I am disgusted by the boot-lickers, by the helpless people.
3	Sh	I loathe boots, I loathe helpless people.
	J	**e.** *When a person that I like very much got near to me.*
1	F	If a person I like very much approached me.
2	J	If a person I like very much got close to me.
3	D	If someone I like came up to me.
	F	**f.** *I was going to knock down a pedestrian with my car.*
1	A	I was trying to push a pedestrian over with my car.
2	F	I was going to knock over a pedestrian with my car.
3	J	I wanted to overturn a pedestrian with my car.
3	Sh	I tried to knock over a female pedestrian with my car.

Table 5.3: Example backtranslations with a loss in emotion connotation correspond to RQ1, with recovered input emotion correspond to RQ2, with a different emotion correspond to RQ3. Input ids are in bold. Input and output emotions are determined by the classifier.

input presents the action as one that the experiencer *"had to"* take. In d. *sadness* replaces *disgust* with the use of a softer expression, such as *"loathe"*. This text further exemplifies that removing a direct emotion word can determine a switch in connotation.

Another reason why the backtranslation in b. is associated to *fear* could be that *"silence"*, in ISEAR, mostly occurs in the description of disruptive, frightening events, similarly to being *"approached"* by strangers (and hence, also the joyful sentence in e. turns into *fear*). There are indicators of emotion changes showing a gender bias (Sun et al., 2019): characterising the subject as a male moves *anger* to *guilt* or *joy* (RQ2 and RQ3, a.), while female characters can elicit an association with *shame* (RQ3, f.).

We notice that localized lexical edits are enough for a classifier to detect different emotion attributes in the generated texts, particularly when the input and output labels can be simultaneously expressed. For instance, as negative emotions, *anger* and *guilt* are more likely to co-occur than *anger* and *joy*, corresponding to output 1 and 3 for the first sentence.

With few exceptions, the texts suggest that transfer can happen without disrupting grammaticality or content within the relatively small top-n lists we considered. Despite this, the overall impression returned by the paraphrases is less satisfactory than it seemed from the quantitative results. Not only did we expect more elaborate paraphrases than what could be obtained by lexical substitution, but hardly do the paraphrases pass a human inspection that checks for the transfer.

The findings derived from the numbers were likely magnified by two interacting factors, i.e., the selection step in the pipeline and the classifier. Our transfer approach was based on comparative and not absolute changes, with a selection strategy that returned the translation hypotheses with the highest increase of a target emotion: only in some cases (e.g., all examples in in Table 5.3) that emotion was the predominant attribute of the texts, i.e., the label receiving the highest probability score by the classifier. Even then, the classifier's decisions are debatable. Besides a few evident mistakes (*disgust* in e.), we believe that the output labels are justifiable (*shame* and *guilt* in examples c. and f.), like humans' subjective views in the previous chapters, but they are so only when focusing on a backtranslation per se, and not side-by-side with the input text (examples 1 vs. 3 in f.). Comparing the input and output labels while looking at the linguistic transformations that led to classification differences convinces us that the transfers were not meaningful. A case in point is the switch from *anger* to *guilt* in b., due to the replacement of the conjunction with a comma, of "*dumb*" with "*silent*" and "*no story*" with "*nothing*".

These conclusions align with past work in several respects. De Mattei et al. (2020) noticed that humans are not as sensitive to changes in style as machines are. Likewise, what we consider marginal emotion changes between input and output are sufficient signals for the classifier to predict different emotion labels. We still need a better understanding of the contrast between our insights and those of Mohammad et al. (2016b), who argued that altering polarity hampers human's ability to determine the original sentiment of the text but does not mislead automatic predictions.

The lexical substitutions allow us to compare our results to those of Helbig et al. (2020) and confirm that an explicit style–content disentanglement (to which we did not aim, but which was often produced by the translation models) is suboptimal for the transfer of emotions. It could have been avoided by endowing the pipeline with an additional re-ranking term, to promote lexical diversity between input and output. However, we intended to test the effectiveness of a MT tool with minimal intervention on its workflow; plus, diversity was already taken care of by the sampling method used for decoding.

On a positive note, these small lexical changes do not flip the meaning of words that could be deemed style-bearing (e.g., "*disgusted*" in example d.). This is encouraging, because markers of emotion attributes seem to carry a substantial part of the meaning of texts. Attempting to substitute those and to maintain the text meaning untouched is a remarkable challenge to the compositionality principle, according to which the meaning of a complex utterance is given by that of its constituents (Frege, 1892). Besides, studying evaluative "styles" like emotions and sentiment by assuming that they reside in individual words is a reduction of their complexity (one which I argue against in the next chapter). Such an outcome puts us aside from Helbig et al. (2020) and other style transfer studies operating explicit disentanglements (e.g., Li et al., 2018; Madaan et al., 2020). They aim at the editing of style markers, precisely, and in consequence struggle to satisfy the content preservation objective (particularly with sentiment).

Style-bearing words are evident in explicit emotion expressions (d.) but not in implicit ones (e.g., f.). Moving from a negative to a positive emotion while keeping content unaltered was indeed more difficult for the former type of text than for the latter. However, to pick up on

Helbig et al.'s comment regarding the varying difficulty of the transfer for different emotion expressions, we found no evidence that specific texts consistently facilitate the task (e.g., for all input–target attributes), which suggest that different transfer treatments might be more or less appropriate in different cases.

Therefore, our manual investigation supports that striking a balance between the three style transfer desiderata is problematic. Meaning preservation came at the cost of only minimal changes in style, and automatic metrics misrepresented the quality of the output. As a solution, the current literature advises conducting extensive human-assisted evaluations (Briakou et al., 2021a,b), which would be beneficial also for this study, to fully understand if annotators look at the style of text to decide on its emotion. We refrain from doing that, given the tepid variations observed in our paraphrases, but we use these to reflect on how to move forward, as I detail next.

6 Conclusion: A New Agenda for (Emotion) Style Transfer

Are emotions in the style of texts? This chapter addressed the question through style transfer, a natural language generation task aimed at rewriting existing texts, specifically to create paraphrases that exhibit some desired stylistic attributes. The goal was to understand if emotions are attributes of this type, and if one can adjust the text affective connotation in the same way as, e.g., a formal letter can be repurposed in an informal way, a literal message can be conveyed with the use of figures of speech, a novel can be edited by mimicking the style of some well-known authors. I ventured the transfer goal with the tools of MT and I answered the question in three steps.

RQ1 raised the issue if automatic translation retains the emotion of texts. We found that a state-of-the-art NMT system tones down emotion connotations, thus presenting a problem for the development of affect-aware MT products, for cross-lingual research based on the translation of existing data, and for communication aided by MT.

RQ2 investigated if the problem of emotion loss can be amended without re-training the models, but by incorporating them with an emotion-informed subselection of translation candidates. Results showed that such a straightforward strategy helps preserve emotions and does not adversely affect translation accuracy.

It should be noted that we relied on a backtranslation pipeline instead of a real-world translation scenario. This motivates an important next step, such as the development of a classifier that estimates emotion probability distributions in multiple languages in a comparable manner – the issue remains, though, how to measure their comparability and how to optimize that measure. Moreover, the analysis relied on a single NMT system, namely FAIRSEQ. This tool is representative of a range of systems. Still, our analysis can be extended to other models and other target languages.

RQ3 shifted attention on style transfer. The same post-processing methodology used in RQ2 was leveraged to induce emotion variability rather than recovering the original one.

The re-ranking proved promising to search differently-connotated paraphrases, but in its current state, it does not represent a sufficiently robust style transfer method to deploy in practical application scenarios, from online communication (e.g., as an assistant) to studies within NLP (e.g., for data augmentation). The little variations it induced, which were too elemental for a task that should mimic language creativity and situatedness, suggest that an emotion objective could improve our pipeline. A viable approach to keep testing the feasibility of emotion transfer is to train translation models with an emotion loss term next to the general-purpose objectives of fluency and adequacy. Tebbifakhr et al. (2019) proposed an optimization strategy for NMT that uses feedback from a downstream task (e.g., polarity classification) to generate translation candidates appropriate for that task. An emotion-based objective could be exploited following their trails, for example with the goal to optimize different translation models that sample hypotheses reflecting *anger*, or *disgust*, or *fear*, etc.: given an input text, the use of one or the other system would be determined by the target emotion. Incorporated in our setup, such models would promote the desired affective load starting from the pivot language already. A comparable MT framework with emotion-aware models could be achieved by training systems based on fluency and adequacy (alone) on parallel data that express different emotions in the source and target sides – however, to date, large parallel datasets labeled with emotions are missing in the field.

By revising the transfer outputs in the light of my initial assumption ("Style is all that can be transferred by style transfer algorithms"), one could conclude that emotion is not a style. At least, not one that can be transferred effortlessly, without touching the meaning of texts as well. Still, the implications of these findings are deeper than a prompt removal of emotions from the sphere of style: they regard style per se, the attempt to transfer emotions, and the whole field of style transfer in general. As a guide to the following discussion, where I stress out the gaps left open by this study and ways to fill them in, I will refer back to the hierarchy depicted in Figure 2.5.

6.1 Lessons on Styles

O, Pushkin, for my stratagem:
I traveled down your secret stem,
And reached the root, and fed upon it;
Then, in a language newly learned,
I grew another stalk and turned
Your stanza patterned on a sonnet,
Into my honest roadside prose–
All thorn, but cousin to your rose.

On Tanslating Eugene Onegin, The New Yorker (1955 January 8)
VLADIMIR NABOKOV

Here and in chapter Chapter 2 I indistinctly called "style" both linguistic variations (e.g., formality) and aspects that underlie them (gender is linked to, but is not, style). I gave a loose characterization of it, adapting one that is established among linguists (Bell, 1984), as something that propagates in text while correlating to factors external to language, like features that develop or are stable within an individual (e.g., emotion, gender, personality) as well as external conditions (the communicative goal affecting registers, formality, etc.).

Hence, like other works in the field, I did not undertake the challenge to define "style" myself. "Operating at all linguistic levels (e.g., lexicology, syntax, text linguistics, and intonation) [...] style may be regarded as a choice of linguistic means; as deviation from a norm; as recurrence of linguistic forms; and as comparison" (Mukherjee, 2005).

What a "deviation from a norm is" comes difficult to pinpoint when dealing with emotions, because they can be recognized in normed, standardized texts (we have seen that in the descriptions of Chapter 3). Next, the "recurrence of linguistic forms" suggests that style emerges from a repetition of patterns over time, ideally observed within the same author, while the study I presented did not dispose of abundant production from each writer. This identifies a gap for an agenda of future research.

I followed the style transfer habit to focus on an individual "style"[4], but there are reasons to conjecture that changing one style implies changing others. V. Nabokov's words in the epigraph above capture that style has a multifaceted nature. They summarize how even transformations that aim at avoiding a distortion of style (i.e, translation) in the end dismantle it and re-write the original text in a "language newly learned". *Age*, *politeness*, *register*, *genre*, and all other styles that style transfer observes in watertight compartments compose a phenomenon that, when modified, is changed in fact in different respects, and as such it must be regarded: style is a multi-dimensional concept, and styles cannot always be told apart from one another. Future work could go in this direction.

The involvement of more than one style in a transfer operation directly concerns the linguistic realization of emotions which relate, for instance, to the utterers' age (Pennebaker and Stone, 2003) – a style under the branch of *persona* in the hierarchy of Figure 2.5. Informative on the matter are several studies in NLP that do not revolve around transfer but focus on the link between affective states (e.g., *emotion state*) and figurative language (i.e., *literality*) (Riloff et al., 2013; Mohammad et al., 2016a; Felt and Riloff, 2020).

The possibility that style transfer does actually style*s* transfer can be substantiated also for other linguistic properties, which are strictly interconnected in the hierarchy and could fit different spots in it. As an example, we have put *literature* under *conventional genres*, but targeting a shift in the literary style of an author touches upon her stable writing patterns (more leaning towards *persona*). Style contaminations are visible in the input texts per se, for instance within the *intended* styles (e.g., a formal register can be instantiated in a genre of a prose as well as in a sonnet), and between them and *unintended* subsets (e.g., one can write a poem while being romantic, and a certain cultural background

[4] An exception is Kang et al. (2019), who transferred multiple *persona* styles in conjunction (e.g., *education* and *gender*).

can emerge while being more or less polite).

At the same time, only some combinations of stylistic attributes are acceptable, as pointed out in an investigation of style inter-dependence (Kang and Hovy, 2021). For instance, the presence of impoliteness in technical writings might be paradoxical. Similar observations hold for emotions. Anger might be conveyed impolitely, but the co-presence of the latter attribute with shame is more difficult to imagine; some emotions might appear with specific types of registers and genres and not others (e.g., hardly do emotions appear in technical writings). Reflecting upon the understanding that not all attributes can be given to all semantics calls into cause another important idea (on which I expand later): style is intertwined with meaning.

6.2 Lessons on Emotion Style Transfer

While the qualitative results for RQ3 invalidated the stylistic nature of emotions in language, this finding is to be taken with some caution and delimited to the operational framework I chose. There are reasons to still see the question unsettled. For one thing, (manually found) examples persuade that an acceptable balance between the three criteria can be achieved, *at least for some expressions* (e.g., relief: *"The fire that burned the wood was extinguished before it hit the houses."* $\rightarrow$ anger/sadness: *"The houses are safe, but the forest is devastated."*).

Secondly, one could argue that the method I introduced is not powerful enough for the transfer, because the affective variation induced by the post-processing is unfairly outplayed by the fluency and adequacy for which the top-n hypotheses were produced. I also neglected the peculiarities of the "style" in question, assuming that the vanishing of styles in MT could be exploited to transfer emotions as it has been for different linguistic properties; but each style might require a robust understanding in itself, as a pre-requisite for the transfer systems' choice and success. Methods that are acceptable to alter, e.g., sentiment, like retrieval-based frameworks, might miss the mark for other styles (Yamshchikov et al., 2019). There is indeed a glaring difference between emotion attributes and others, in that emotions are not necessarily mutually exclusive. It is even possible that approaches that work well for specific emotion expressions are ill-suited to others, since we are dealing with a phenomenon with tremendous variation in language.

The discrete labels I used account for only part of humans' emotion episodes, and dismiss other important aspects, like the intensity of such experiences (Sonnemans and Frijda, 1994) and the degree of arousal and dominance in the concerned individuals (Mehrabian, 1996). Style transfer could be done in the future based on those dimensions, for example by controlling the degree to which emotion attributes are modified, similar to text generation studies that condition both the emotion and the emotional strength of texts (Ghosh et al., 2017; Goswamy et al., 2020, i.a.). This can make the task more feasible – e.g., the transfer between different emotions might be doable but only for specific levels of intensity.

Lastly, the re-ranking of backtranslations represented an inexpensive solution to perform style transfer without parallel corpora. A to-do for future emotion-aware automatic paraphrasing is the gathering of human-written reformulations. They can provide both a supervised scenario (ideal for the systems to learn parallel variations from the data) and a sanity check for the difficulty of the task (to see how the ability we attempt to mimic automatically is put into practice among people).

6.3 Lessons on Style Transfer

The results of RQ3 also permit us to reflect on the tool (i.e., style transfer) that I employed to understand if the equation emotion=style is defendable. In fact, they suggest some practical adjustments for the field, specifically to relax some conditions under which it currently operates.

Reducing the Space of Styles. The expectation that a text can be paraphrased with any attribute corresponds to presuming the existence of groups of "semantic equivalence" that subsume texts with similar meaning but different stylistic attributes. This view has an evident yet neglected consequence for the field: if collocating an attribute under a specific meaning seems unfeasible, then such an attribute cannot define a goal for style transfer. That holds for emotions but not only them. The paraphrases in this study raised issues similar to those observed in others that sparked energetic debates (Tikhonov and Yamshchikov, 2018), particularly evaluative styles like sentiment, whose transfer comes at the expense of losing content, contradicting the meaning–style independence assumption.

Thus, emotion casts doubt on the possibility of addressing style transfer with any feature of text that can be shifted along some dimensions, and that appears to tie in with some extra-propositional content of texts. **If** it is not a style, it is so in the same way as other linguistic features, and the field should reduce the space of styles it uses. In fact, the multiplicity and diversity of styles in the existing publications (e.g., *gender*, *politeness*) invites us to reconsider if different works discuss the same phenomenon.

Reducing the Space of Evaluation Criteria. As an alternative, evaluation approaches can be refined for said styles. A style X might well be part of "all that can be transferred", but what it means for *that specific X* to be transferred needs clarification. The literature already revealed that the available metrics for content preservation, naturalness and transfer strength are inadequate. For instance, n-gram overlap measures like BLEU (included in the transfer system of this study) both disfavors diversity in the output and does not weight style-relevant edits over the others (Krishna et al., 2020). However, the problem might reside upstream: the transfer criteria arguably generalize across styles, each of which has particular characteristics. Is a successful system for the transfer of emotions supposed to maintain meaning as much as a politeness-conditioned system? Precisely because different styles have different linguistic realizations, it seems somewhat unreasonable to expect that the systems addressing them operate under the same constraints. Transfer, meaning, and grammaticality may be variously reached for each style, making it more urgent to ask *to what extent can a method changing the emotion/sentiment of a text retain its semantics?*, than measuring if it did. An investigation of transfer with respect to individual styles can redefine the task at hand and give a measure of the attainable goals.

Relaxing the Style–Content Independence. Style could be more than what is transferrable. Emotion might well be part of this phenomenon, but under different conditions – for instance with a more flexible operational framework that accounts for the intertwining between style and content, rather than their independence.[5]

Content is information predictive of a future (e.g., what word comes next?), while style is additional information prior to generation. It is grounded in reality, in the human experience (e.g., gender, ethnicity), and ultimately, in the reasons that push speakers to communicate and that current machines (struggling to transfer) do not have. Because style precedes generation, it can determine content. It is sufficient to think of scientific language, or of children's books, to see that style

[5] In this regard, Kang et al. (2019) found that features used for classifying styles are actually of both a stylistic and a semantic type.

determines how something is said but also what is said in the first place (vice versa for meanings).

All things considered, this chapter endorses the view that changing any feature treated as style without distorting the semantics of individual sentences is a debatable attempt, and it urges researchers to concede that the style–content independence loosens in some cases.

Should We Agree on Meaning? Defining style is an ambitious goal, and doing the same for meaning comes with an enormous difficulty. One can at least agree on what meaning preservation requires. A tangible step is to ground semantics in modalities other than text. The classes of semantic equivalences mentioned above can be defined, e.g., in a space of vision. Input and output texts would be compared to their potential associations with images, such that the preservation of content c in the paraphrase t' of t is not given by the approximate equivalence of $p(c \mid t) \approx p(c \mid t')$ but by $p(image \mid t) \approx p(image \mid t')$, as a way of comparing their common association to sceneries or background frames of reference.

Semantic frames are what I use to model meaning in the next chapter, but remaining within the modality of text and leaving the transfer goal behind.

Chapter 6

Characterizing Emotions in Frame Semantics

The finding that emotion in language is not a style, at least for a certain definition of "style", brings us to consider meaning. Dictionary-based approaches to emotion recognition posit that affective phenomena are entangled with semantics and treat individual lexical units as the fundamental emotion cues. However, emotions can be realized in text implicitly, for instance with references to facts (e.g., "*That was the beginning of a long war*"), which suggests that interpreting affective meanings relies on the readers' background knowledge about word relations, an idea that is hardly modeled in computational emotion analysis.

As a remedy, this chapter shifts attention to the "semantics of understanding" of frames, a theory for the description of predicate-argument structures. Its overarching question is whether and to what extent the events that are represented by frames possess an emotion meaning. I describe a corpus-based correspondence analysis that glues a theory of emotions from psychology (appraisals) with a theory of semantics (frames). Quantitative and qualitative results show that substantial groups of frames have an emotional import, and that they capture several properties of emotions purported by appraisal theories.

Highlights

Emotion analysis can profit from the event-based perspective of frame semantics.

Emotion meanings reside in word relations.

More frames than those attested as emotional in FrameNet have an emotional side.

Emotional frames capture appraisal-based emotion properties.

1 The Need for an Emotion-Based Event Semantics

Understanding events is fundamental to discuss emotions, which not only emerge in real life as responses to salient circumstances, but they are also communicated via verbal realizations corresponding to descriptions of facts and states of affairs. Writers can describe a stimulus event (e.g., "*my grandad died*", "*my team won the match*", which likely spark sadness and joy), or their reaction to it (e.g., "*I cried*", "*I smiled*"), trusting that an emotion interpretation of their production will be drawn by the readers via pragmatic inference (Grice, 1975). Chapter 3 showed indeed that emotions can be inferred from text when there is not much pertaining solely to the linguistic material that discloses an affective intention or an affective tone.

How can emotions be associated to these factual statements? While psychology calls into question empathy and affective role taking (Mehrabian and Epstein, 1972; Eisenberg and Miller, 1987; Omdahl, 1995), natural language processing links emotion decoding more directly to world knowledge. Its foundational observation is that words possess specific affective meanings in the collective imagination (Clore et al., 1987) – e.g., *die*: sadness, *win*: joy, *ghost*: fear. A way of quantifying such meanings in text is to represent them with measures of word-to-emotion association that are stored inside dictionaries (Strapparava and Valitutti, 2004; Mohammad and Turney, 2013).

Dictionaries thus assume that individual words are the crucial, emotion-revealing linguistic units. This view is practically useful, but it neglects that the affective substance of texts is often determined by *relations* between words, both by their contextual role in a linguistic structure, and by their semantic functionality (synonymy, antonymy, etc.). For instance, the prototypical sadness-oriented meaning of "*die*" can be capsized by accompanying this word with certain others (e.g., "*my archenemy died*"); similarly, in the sentence "*my opponent won the match*", the verb "*win*" can be argued to diverge from its usual joyful connotation. Dictionaries do not account for the linguistic context in which lexical units appear, and therefore, they fall short in using the paradigmatic and syntagmatic bits of information that allow people to infer emotions. Approaches that embed emotion meanings into latent vector spaces (Felbo et al., 2017; Li et al., 2017, i.a.) capture this contextual information, but they are less transparent to investigation than lexical-based methods.

Frame semantics (Fillmore, 1982), which to the best of my knowedge has never found its way into the computational study of emotions, checks all marks. Its "semantics of understanding" or U-semantics (Fillmore, 1985) would consider "*my grandad died*" and "*my archenemy died*" as revealing different types of relationship between the writer and the subject of discourse. Put differently, frame semantics offers a perspective on text interpretation that leverages people's knowledge about how the world is organized. It proposes a formalism (i.e., frames, set by the Berkeley project in the FrameNet database (Baker et al., 1998)) that grasps the meaning of textual chunks larger than individual words, such as whole predicate-argument structures.

These aspects alone suggest that frames bear a potential value to study emotions, but there is an even more fundamental affinity between them: since frames are abstractions of real-life situations, they have a focus on events like appraisal theories of emotions (cf. Page 21). Descriptions of emotion-triggering stimuli (e.g., "*my team won the match*") can thus be represented by frames. It is reasonable to deem also other emotion expressions as reporting (frame-evoking) events, from the very occurrence of an affective experience (e.g., "*I'm happy*") to a related response (e.g., "*I'm all steamed up!*"). What is more, frames contain information about semantic roles (e.g., what happened and who took part in it). Therefore, they lend themselves well to model the components involved in an emotion episode, such as the emotion-arising event, its experiencers, and relations among them, all of which have linguistic realizations. In this sense, work in structured emotion analysis that aims at identifying "who feels what and why" (Oberländer et al., 2020; Wei et al., 2020) constitutes a domain-specific instantiation of the general

U-semantics question of “who does what to whom”.

The conceptual similarities between frames and emotions give a compelling reason to put into scrutiny their relationship, as a way to complement the investigation of emotions in the style of text with one that looks for them in textual meanings.

2 Emotions: A Dimension of Meaning?

As introduced in the Background of the thesis (see Page 65), there exist some frames that FrameNet dedicates to emotions (e.g., EMOTION_DIRECTED and EMOTION_OF_MENTAL_ACTIVITY). They compose a small emotion nucleus, and by navigating relations within the FrameNet vocabulary, an affective load can be reconstructed also for other frames. For instance, FLEEING can be related to the FEAR frame via the USE relation (Ruppenhofer, 2018).

For the majority of them, a link to emotions (or to a lack of emotions) is not specified. Even the exact vocabulary belonging to the emotion domain is not spelled out, partly because FrameNet is a database under constant development, and partly because emotional meanings are only one type of inference that can be made from frames – representing all of them would be unfeasible for the FrameNet curators. Yet, the finding that emotions can be inferred from factual texts (cf. Chapter 3) leads to believe that much of the background knowledge modeled by frames has an affective side. Possibly, that happens with more frames than those reported in said emotional nucleus.

Identifying emotion-bearing frames systematically would be relevant from two complementary perspectives. Frame semanticists might use the (eventual) understanding of emotions as an underlying component of the meaning of frames. For researchers in emotion analysis, FrameNet could represent a suitable tool to tackle the emotional import of sentences. It could guide the field toward better automatic text interpretations and, in the context of this thesis, provide insights on the linguistic level at which emotions come to actualize (e.g., in the relation between words and not in isolated words).

Goal of the Study. Like dictionary-based approaches, this chapter studies meanings, but the meaning captured by frames. It asks: Are FrameNet frames associated with (discrete) emotions?

Because we strive to understand the affective side of semantics, we take a relaxed approach to discrete models from psychology. We conflate diverse emotion classes under the umbrella phenomenon of *emotionality*[1], i.e., whether a text has an emotion content, irrespective of what it is. Therefore, to refine the research question, we ask: Are FrameNet frames associated with emotionality? (Many of them are, as it emerges from a quantitative analysis.) Our hypothesis is that FrameNet can be divided into three classes: a first group of neutral frames that have no emotional import (e.g., STORING); second, a group of frames which are evoked by unambiguously emotional predicates (e.g., EMOTION_DIRECTED, FEAR); and lastly, following the theme of the implicitness of emotions from Chapter 3, a group including frames that express factual but emotional events (e.g., DEATH). (With a qualitative analysis, we also confirm this hypothesis.)

We answer the research question empirically, endowed with a large corpus of contemporary American English. The corpus contains varied types of emotion realizations, including but not limited to factual statements and implicit expressions. We perform two strands of automatic annotation on it.

The chapter starts by introducing our experimental setting (Section 3). Section 4 presents our main contribution, which also elaborates on the tripartition of the FrameNet emotion vocabulary. Section 5 follows up with a discussion of the implications of our findings. I conclude with an outline of the limitations of the present work and I indicate ways to expand it (Section 6).

3 Methods

Studying if the emotionality of texts is mirrored in the frames that the texts evoke, and if this, in return, can help making sense of FrameNet is close to NLP tasks that exploit human-curated resources in order to improve embeddings. The rationale is that word representations and prior knowledge about language, as it is contained in lexicons, are best used in combination because each of them captures some dimensions of meaning that the other overlooks. For instance, Faruqui et al. (2015) ensured strong conceptual alignment between semantic lexicons and word vectors, and found that this increased the quality of the latter. Also successive works (e.g., Kuznetsov and Gurevych, 2018) used lexical resources to boost the performance of data-driven learning methods. We move at the interface between these two methodologies as well, but

[1] I henceforth use "emotion" and "emotionality" interchangeably, within the chapter.

with a reversed perspective: emotionality is a dimension of meaning that can be obtained from approaches based on word embeddings; with the latter, we understand if it is also grasped by FrameNet, and eventually remedy the limits of the database in this regard.

Below, I describe the textual source we used, as well as our methodology based on prior knowledge and data-driven strategies, highlighting the challenges it poses.

3.1 Data

In order to generalize our empirical observations, we aim at collecting texts that showcase a variety of linguistic realizations of emotions, and that evoke frames across both emotion-bearing and neutral expressions. Obtaining the two types of information is not straightforward. There are no resources for emotion analysis labeled with frame semantics information, except for the dataset by Ghazi et al. (2015), which is limited in size and only includes emotion-bearing texts. On the other hand, corpora for frame-based semantic parsing do not contain emotion annotations – at least not for the vast majority of frames, as discussed.

Hence, we base our study on an unlabeled corpus, the 2020 version of COCA (Davies, 2015), which is much larger than any existing resource for emotion analysis.[2] Like in Chapter 4, we consider all domains that COCA comprises[3], except for academic texts as they have an arguably unaffective language; we split textual paragraphs into sentences, and we exclude sentences containing words that are masked for copyright reasons. We further filter out all those with less than 3 tokens (tokenization performed with the python library *nltk*). The data so preprocessed comprises ∼44M sentences and ∼536M tokens.

3.2 Bridging Data-driven Learning to Semantic Resources

To obtain frames and emotion information, we bypass the use of human annotation which would not scale easily to the amount of data we aim to study. We resort instead to an automatic procedure, adopting a two-step methodology (illustrated in Figure 6.1): first, texts are associated

[2] An overview of resources for computational emotion analysis can be found in Bostan and Klinger (2018).

[3] Blogs, magazines, newspapers, spoken interactions, fiction, TV and movie subtitles, webpages.

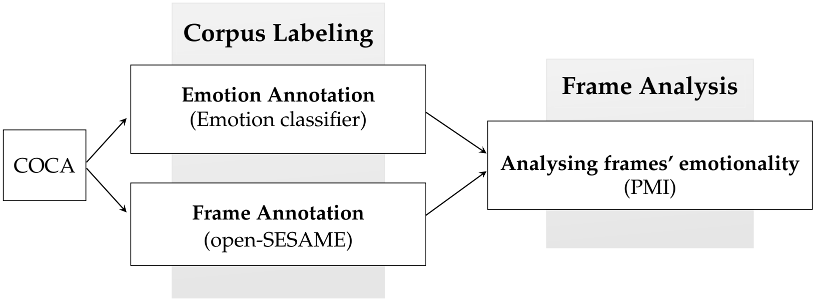


Figure 6.1: Our two-step experimental setting. Corpus Labeling: automatic annotation of sentences extracted from the corpus of contemporary American English with emotions and frames, separately, with the emotion classifier being evaluated on a subset of the corpus previously annotated by human judges, and the semantic role labeler evaluated on a subset of MASC as out-of-domain data. Frame Analysis: the two strands of annotations are brought together to observe the association of frames and emotionality via PMI.

to emotion labels with an emotion classifier, and to frames with a SRL tool; second, we carry out a corpus-based correlation analysis where the association between the two annotation sides is quantified and interpreted.

This type of approach is common to data-driven information extraction lines of research which require no human intervention, such as the task of open information extraction (Etzioni et al., 2008), as well as to distant reading, i.e., the application of computational and statistical techniques in the field of digital humanities, aimed at uncovering global patterns in texts (Jänicke et al., 2015). Still, we acknowledge that both the emotion classifier and the parser might make mistakes. The data we study was not collected for the sake of computational emotion analysis nor semantic role labeling, and might differ in tone, topics and linguistic structures from the resources on which our automatic annotators were trained. As a matter of fact, the generalization capabilities of FrameNet-based parsers have been put into question by Hartmann et al. (2017), who found that a state-of-the-art system for SRL loses 16 percentage F_1 points when evaluated against out-of-domain data. This issue exacerbates when considering emotions. Bostan and Klinger (2018) showed that systems for emotion detection tested out of domain suffer from performance drops as heavy as $\approx$.70 in F_1 score.

To address this potential problem, we adopt experimental design strategies targeted at reducing the presence of random effects in our findings. One, already mentioned, is to employ a corpus with a considerable number of datapoints. Second, we carry out the annotation with artificial neural network-based technologies that have shown the ability to generalize well over unseen sentences and predicates. Third, we evaluate the emotion classifier against a manually-annotated sample of our texts as an additional check of its reliability, and we do the same for the semantic role labeler using out-of-domain data. Lastly, we conduct a statistical analysis that makes the observed frame-emotion associations unlikely due to chance. Overall, our findings should be interpreted in relation to the systems that we employ, but on the other hand, they do provide evidence to learn about the bond between frames and emotionality.

We now proceed to describe the instantiation of all components involved in our investigation shown in Figure 6.1.

3.2.1 Corpus Labeling

As a first step, we associate texts to emotion labels and to frame-related information. The systems used here are trained separately on different corpora. It is thus necessary to assess their domain independence and get some insight into how well their automatic judgments fit COCA.

Using a resource already labeled by humans (either for frames or emotions) and only performing one strand of research could have be a safer alternative. People's judgments are arguably more reliable than those of a classifier. Yet, other than having magnitudes of data points less than we need, existing resources for affective computing typically focus on a specific type of texts, such as tweets (Mohammad, 2012), tales (Alm et al., 2005) and news headlines (Bostan et al., 2020) while we aim at observing a variety of occurrences of frames (e.g., elicited by different lexical units) and emotion expressions, likely to be found in a mixture of text domains. These resources also tend to have different emotion annotation schemata, which means that the same expression can have a different emotionality depending on the context of utterance. Employing a state-of-the-art classifier specialized in only one domain (i.e., trained on a single resource for emotion analysis) would give no guarantee that the obtained annotations are valid for our data.

Emotion Classifier. We start by gathering various resources for emotion analysis that span textual domains similar to those in COCA, from webpages to literary texts: Grounded-Emotions (Liu et al., 2007), EmoInt (Mohammad and Bravo-Marquez, 2017), TEC (Mohammad, 2012), SSEC (Schuff et al., 2017), enISEAR (described in Chapter 3), ISEAR (Scherer and Wallbott, 1997), Tales (Alm et al., 2005), DailyDialogs (Li et al., 2017), and Emotion-Stimulus (Ghazi et al., 2015). This is the group of resources that in Chapter 4 we refer to as UED. To that, we add GoEmotion (Demszky et al., 2020).

These datasets feature diverse emotion schemata; we make them consistent to our binary setup by mapping their original labels into the *neutral* and *emotional* classes, depending on whether a text was marked as having no emotion, or as having one out of a rich set of alternatives (e.g., *joy*, *fear*, *disgust*, *hope*, *surprise*, *guilt*).

Instead of extracting our test set from this data, we use the portion of COCA introduced in Chapter 4. The sample contains 700 texts labeled at the sentence level by three in-lab raters. They are balanced across the domains that we consider, and their annotation encompasses the same binary categories of our concern: the neutral label corresponds to the absence of any emotion content, the emotional class represents the quality of a sentence that has an emotion as a central component of their meaning or that describes an event, a concept or a state of affairs to which the annotators would personally associate an emotion (without aiming to reconstruct the one felt by the writers).

Next, we train a number of models on the concatenation of the selected (training) resources: we fine-tune BERT (Devlin et al., 2019) base-case models, adding a classification layer that outputs the labels *emotion* or *neutral*, similar to what we did in Chapter 4. Different models are obtained by varying the data on which they learn the classification task. Our objective is to identify a subset of training resources that yields a classifier capable of reliably judging out-of-domain data (i.e., COCA). Hence, we evaluate each model on the manually annotated COCA sample, with the majority vote determining the ground truth, and we pick the model that performs the best on this test set to annotate the rest of the corpus.

Model selection is shown in Figure 6.2. Classifiers are plotted as dots in the figure, numbers on the x axis correspond to how many datasets are removed at each successive step. Recursive data elimination proceeds as a backward search, in the following way. Initially, we train a classifier on all gathered corpora (“D” in the figure, F_1 = .59); from these resources, we pull out each dataset separately (“D -1”), and observe that the ablation of DailyDialogs is the most beneficial (F_1 increases to .65); we move on to the next ablation step and keep

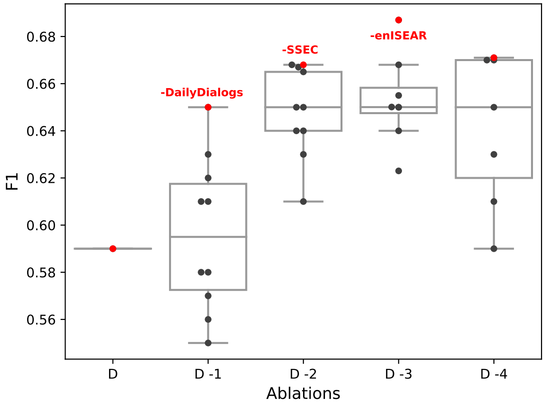


Figure 6.2: Model Selection: weighted F_1 scores (y axis) of the models evaluated against the annotated COCA sample. We recursively ablate datasets from the training set that yields the best model at the previous step (x axis). Dots are classifiers obtained with an ablation; the red ones indicate the best performing model: from all datasets (D), we remove each separately ("D −1"); from the set on which we obtained the best model (red dot "−DailyDialogs"), we again we remove each dataset, one at a time, thus training the next models on a collection with two datasets less than D (i.e., "D −2"); and so on.

using the data that yielded the best performance. From that, we ablate each remaining dataset (i.e., "D −2"): now, the results reached upon removal of SSEC surpass the previously best classifier. We repeat this procedure and reach an upper bound F_1 score.

From the total of 35 trained models, the most competitive one is obtained when removing DailyDialogs, SSEC, and enISEAR (F_1 = .69 with "D −3", which outperforms the best model in "D −4", F_1 = .67). We use that to annotate COCA.

It might be noticed that the performance of our BERT-based classifier on COCA is not state-of-the-art in emotion classification. However, systems for emotion detection that work well on existing labeled resources (such as the kin classifier described in Chapter 4) might not be equally well-performing on COCA. We varied the model architecture and noticed that a model that achieved better results on in-domain data suffered from major performance loss when evaluated on the manually-annotated subsample of COCA (further details, training hyperparameters, and discussion in Appendix 4.1).

Semantic Role Labeler. Models and corpora for semantic role labeling

	Frame Id		
	P	R	F_1
FrameNet 1.7	.85	.85	.85
Masc	.78	.78	.78

Table 6.1: Evaluation of the SRL tool provided by Swayamdipta et al. (2017) against FrameNet data and MASC frame-annotated data.

are scarcer than emotion-centered ones. Here, we require a semantic role labeler which, given a sentence, identifies the set of FrameNet frames that are evoked by each of the predicates, as well as the corresponding predicate arguments. To this end, we use open-SESAME[4], a freely available interpreter for SRL with state-of-the-art performance, which was developed by Swayamdipta et al. (2017). Their models are based on bidirectional recurrent neural networks and semi-Markov conditional random fields (Sarawagi and Cohen, 2004). We re-train the implementation they provide[5], using the sentences from the FrameNet release 1.7 (7340 for training, 387 for dev, and 2420 for testing).

We evaluate the semantic role labeler on the FrameNet test set as in-domain data, as well as on external data. For that, we use 695 sentences (516 of which are frame-evoking) coming from MASC[6] (Ide et al., 2010), a subset of the Open American National Corpus that provides useful annotations for frame-based SRL. The texts in MASC include emails, essays, fiction, and spoken transcripts. Using this resource as a benchmark illustrates how the SRL tool performs on linguistic expressions similar to those found in COCA.

Precision, recall, and micro-averaged F_1 are reported across both test sets in Table 6.1 for the frame identification task (Frame Id). We obtain these results using the script by Swayamdipta et al. (2017), based on the full-text FrameNet annotation and following the standard evaluation practice set by the SemEval 2007 (Baker et al., 2007).

A drop in performance can be noticed when moving to out-of-domain data (from F_1 = .85 to F_1 = .78). Conjecturing that this is partially due to an increase in sentence length (avg. for the FrameNet test = 16.5 tokens, for the MASC test = 23.4 tokens) and in the average number of frames per sentence (2.8 for FrameNet, 6.45 for MASC), we

[4] `https://github.com/Noahs-ARK/open-sesame`.

[5] Training hyperparameters as chosen in Swayamdipta et al. (2017).

[6] `https://www.anc.org/MASC/download/MASC-1.0.3.tgz`.

deem these numbers as evidence that the considered SRL system can be used to proceed with the annotation.

3.2.2 Frame Analysis: Investigating Emotionality in the Frame Structure

Once COCA is labeled with emotion labels and frames, we can address our research question. For that, we estimate the degree to which the two phenomena are associated. This step requires an appropriate alignment strategy, as the labels obtained so far have a difference in granularity: emotions are associated to entire sentences, while the output of the frame parser relates to tokens.

It would be reasonable to weight the alignment between a frame and an emotion label of a sentence based on whether (and how many) other frames appear in it. At the same time, not all weighting schemes work equally well for all tasks (Buckley, 1993; Pekar et al., 2004; Ushio et al., 2021). Selecting one that fits our study would pose an additional impasse. Our solution is to consider each frame in a sentence as having a separate and full-fledged alignment with the emotion label, without any weight specification – a design choice that is not only transparent but also comparable to related work (e.g., aspect-based summarization in sentiment analysis (Hu and Liu, 2004), where multiple aspects identified at the sub-sentence level are grouped under the same sentiment label).

To disclose patterns of frames occurring with specific emotion labels, we then use pointwise mutual information (PMI) (Church and Hanks, 1990). This information-theoretic measure quantifies the dependence between the values that two discrete random variables can take, and accounts for their chance co-occurrence. More specifically, PMI compares the probability of observing two variables together, against that of observing them independently, or by chance. In our case, the variables are the output labels of the automatic annotation procedure from the corpus labeling step. For each pair (f, e) consisting of a frame and an emotion label, we estimate PMI as the number of times that such frame and emotion label co-occur in the entire corpus, divided by the product of their individual frequencies. Formally, for each f and e, we compute

$$\mathrm{PMI}(\mathrm{f;e}) = \log_2 \frac{\mathrm{p(e, f)}}{\mathrm{p(e)p(f)}} = \log_2 \frac{\mathrm{p(e \mid f)}}{\mathrm{p(e)}}.$$

	Emotional	Neutral
Sentences with Frames	19.717.813	16.092.214
Sentences w/o Frames	4.194.783	2.141.299
Number of Frames	75.889.290	57.517.465

Table 6.2: Outcome of Corpus Labeling. Number of sentences associated to the emotion and neutral labels, both with frames and evoking no frames, and number of frames.

As already mentioned, the number of extracted (frame f, emotion label e) pairs varies from sentence to sentence, depending on whether one or many frames are evoked.

PMI does not have predefined bounds. Positive values indicate that a frame and an emotion connotation are semantically associated: they appear together more than one could expect by considering the two events independently. PMI = 0 represents no reduction of uncertainty in predicting the outcome of one event (*emotional/neutral*) having observed the other (a specific frame) – i.e., there is no dependency between the two. Lastly, negative values indicate that f co-occurs with the considered e with less than chance expectancy and therefore is associated more with the opposite emotion label.

4 Emotion-Frame Associations

The processing steps detailed in Section 3.2.1 provided two independent layers of annotation for the same texts. Statistics describing this outcome are reported in Table 6.2. The emotion labeling module resulted in ≈23M sentences labeled as emotional and ≈18M as neutral. From this total, ≈6M sentences (i.e., ≈4M emotional and ≈2M neutral, row "Sents. w/o Frames") were not associated to any frame by the SRL system. In our analysis, we do not consider these frameless sentences, which typically consist of short texts like *"That's what it was"* and *"-No, it's not a guy"*. For all others (row "Sents. with Frames"), the role labeler identified 133M frames, specifically in the 76M emotional sentences (816 unique frames) and 57M (818 unique ones) in the neutral counterparts, with an average of 3.7 frames per sentence. Only a few frames are evoked in one group of texts but never in the other. DYING is only identified in emotional texts, while ATTENTION_GETTING,

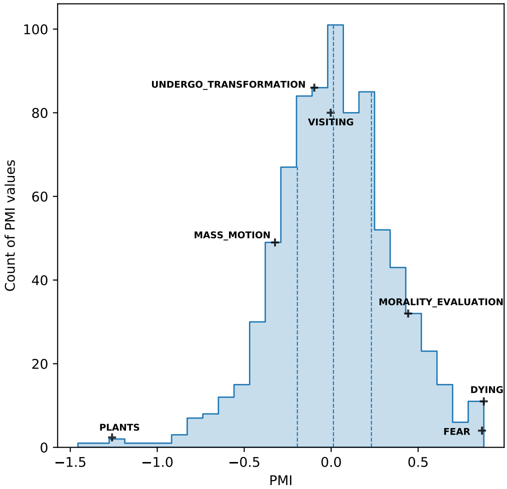


Figure 6.3: Histogram of PMI(f;emotional). Dashed lines: beginning of 2^{nd}, 3^{rd} and 4^{th} quartiles (respectively, PMI = −0.19, 0.01, 0.23).

SPECIFIC_INDIVIDUAL and ADDICTION are only associated to neutral texts.

To discuss the emotion-frame association, we analyze the PMI between all identified frames and emotionality. In this binary classification setup, the distributions given by PMI(f;emotional) and PMI(f;neutral) are essentially symmetric. Frames which are positively correlated to one label are not to the other. Consider, for instance, the frame MORALITY_EVALUATION illustrated in Figure 6.3: PMI(f;emotional) = .44, while PMI(f;neutral) = −.8. For this reason, we only report the emotional distribution. The whole list of frames and emotionality associations, computed also with respect to the neutral label, is in Appendix, Section 4.2.

As shown in Figure 6.3, PMI values are approximately normally distributed around 0 (there are a few more outliers on the negative than the positive side), with no systematic deviation. This is unsurprising, because the same event can elicit different types of emotion reactions and be represented in language accordingly. Therefore, we take such values as a sign of the reliability of our automatic annotation procedure.

At the same time, the distribution does not provide a clear-cut separation between emotional and neutral frames, which is necessary to discuss emotionality. We could define it in many ways, for instance

Emo. Frame	Text
JUDGMENT_DIRECT_ADDRESS	Oh, thank God, thank God you're not mad at me for pushing you that day.
EMOTIONS_BY_STIMULUS	So glad we're friends .
DISGRACEFUL_SITUATION	This is outright, outrageous, disgraceful, disgusting.
FACIAL_EXPRESSION	Mr. Imperatore smiled at those memories and said he had mended his ways.
CAUSE EMOTION	The whole thing was quite pathetic, really, and insulting to boot.
EXPERIENCER_OBJ	I am surprised the judges bought it.
COMMUNICATION_NOISE	For the first week I cried.
PROTEST	He marched, he organized, he protested, he was gassed, he was beaten, he was jailed.
CONTRITION	Blinking furiously, looking furiously guilty, Jimmy Lowe says, "All's I did – Ziggefoos cuts him off."
EMOTION_DIRECTED	And – and she just made you happy.
Neut. Frame	Text
CHANGE_RESISTANCE	Your child may need braces if his or her teeth are overcrowded, crooked, or overlapping.
CONTROL	Controlling her physical symptoms had occupied most of her time.
STORING	Mark your packages with the date they were placed in the freezer so you can keep track of storage times.
MEASURE_AREA	They burned 665,000 acres; roughly 40% of the statewide total of 1.7 million acres.
Cont.Det. Frame	Text
COMMUNICATION_RESPONSE	(N) The **answer** is, you don't, or at least not with career backups. (E) The **answer** would be NO!
GIVE_IMPRESSION	(N) Neither candidate **seemed** to have any awareness of virality . (E)You really **seem** to be exploding with creativity!
POINT_OF__DISPUTE	(N) The **question**, crude as it was, hung in the air . (E) The **issue** is not whether I was a perfect pastor; I was not .

Table 6.3: Top: Examples of emotional frames with sentences in which they appear. Bottom: Examples sentences evoking contextually-determined frames, with (E)/(N) indicating emotional/neutral sentences (text in boldface corresponds to predicates).

using PMI = 0 to decide on what counts as emotional and what not. However, we adopt the quartiles of the PMI distribution because they represent not only a statistical-motivated approach, but also a good balance between the precision and recall of our findings: as opposed to a binary separation, they shield us from considering as emotional some frames with a minimally positive PMI (due, e.g., to bias in the data or mistakes of either automatic labeler); compared to more restrictive cuts (e.g., taking the top 10% of frames as emotional), they facilitate our analysis of what frames other than those known to be emotional are so.

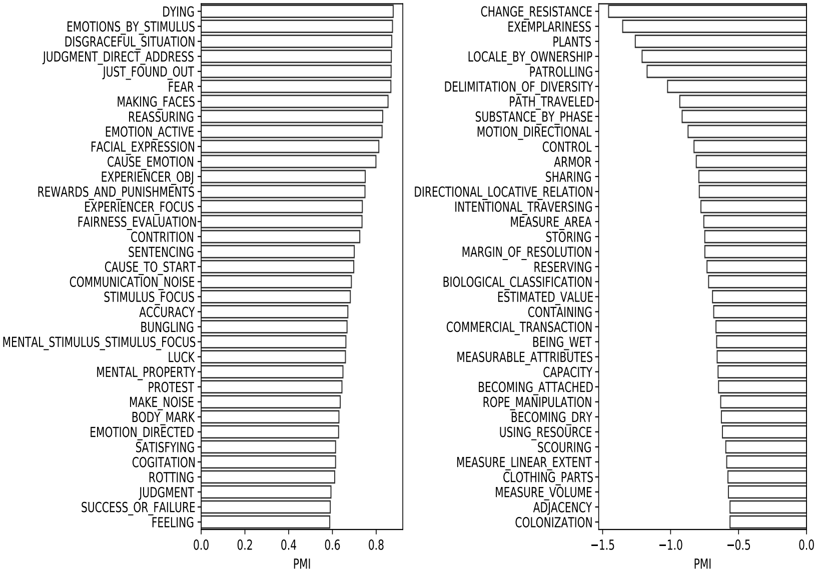
From now on, we deem the top quartile of the distribution to correspond to emotional frames, as it identifies the highest 25% of PMI values positively associated to the label *emotional*. Frames in the bottom quartile will be treated as neutral. Frames that fall within the second and third quartiles are not emotionally- nor neutrally-connotated, and are referred to as "contextually-determined" for reasons discussed in Section 4.2. Both the emotional quartile and the neutral one encompass 204 items, and the contextually determined group comprises 408 frames. Some COCA sentences in which emotional, neutral and contextually-determined frames appear are in Table 6.3.

We next characterize the emotional frames and those which could belong to either label with a qualitative analysis that validates the findings of the PMI procedure.

4.1 Emotional Frames

Figure 6.4 (a) illustrates the 35 highest PMI-valued items. This small subset already reveals how diverse the emotional frames are. Looking at the whole group with the 204 elements, we notice that they range from circumstances of interpersonal communication (e.g., OPINION, REVEAL_SECRET, WARNING) to actions (e.g., RUN_RISK), from internal motives (e.g., WILLINGNESS, RENUNCIATION) to social circumstances (e.g., HOSTILE_ENCOUNTER, PREVARICATION).

A handful of these frames have a clear emotional quality, like FEAR or EMOTION_ACTIVE, and they are treated in FrameNet itself as such. However, for almost all of them, an emotion content is more opaque and ripe for investigation. As an example, FAIRNESS_EVALUATION does not express a clearly affective situation. Therefore, to corroborate the emotionality of the instances in the top quartile of the PMI distribution, we look for qualitative evidence, tapping on emotion theories from psychology. We hypothesize that these frames share a common affect-



(a) Emotion-related frames. **(b)** Neutral frames.

Figure 6.4: (a) The 35 frames with the highest PMI values in the emotional distribution, in comparison to the frames with the lowest values (b). See Table 6.3 for example sentences in which these frames are evoked.

laden ground despite their variety, because many of them well adapt to the diagnostic features of emotions that are explicitly endorsed by theories of appraisal (i.e., presence of an event, event evaluations, concomitant changes).

The idea leading this qualitative analysis was laid out in the communication framework discussed in Chapter 3 (Page 76): an emotion-bearing text can mention any point of the emotion episode, from the stimulus to the component process lived by the experiencer in response to it. Therefore, we cluster the 204 frames into three different semantic groups that map either to the emotion vocabulary of FrameNet, or to appraisal-motivated emotion properties concerning the circumstances that spur an emotion or its occurrence within an individual. The groups are:

(1) unambiguously emotional frames;

(2) frames depicting events/concepts that might cause an emotion

(i.e., emotion stimuli);

(3) frames that capture cognitive evaluations of situations (i.e., the fundamental factor for an emotion to occur) or emotion manifestations (i.e., the factor that appraisal theories see as part of the emotion process).

Regarding group (2), we notice that there is a conceptual divide between frames and events in an appraisal framework. The former are abstractions over many situations. The latter are evaluated with subjective properties contingent on a circumstance, as seen in Chapter 3. The instantiation of appraisal criteria might change with every textual instance where the same frame appears (e.g., the event DEATH can be linked to *others' control* once and to *situational control* another time), leading to specific emotions or to none. Looking at emotionality resolves that gap: what matters in this binary setup is if the situation modeled by the frame is attributed an emotion, and not how that frame is linked to emotionality by virtue of what evaluation. In that sense, emotional frames can be deemed generalizations over different appraisal configurations.

We set guidelines for determining if a frame belongs to one of the three groups. The attribution to one or another is mostly based on the frame definitions in FrameNet. When these turn out uninformative, we make a decision by randomly sampling 20 sentences containing the frame and assessing what aspect of emotions they model.

The outcome of this manual classification is reported in Table 6.4, together with the definition of each group used as a guideline. Numbers preceding the frame identities correspond to their PMI rank. As can be seen from the bottom block of the table ("Idiosyncratic"), 22 items do not fit in any of the three clusters, since they hardly capture an emotion property or an emotion-inducing event. By contrast, most emotional frames fit well into the tripartition that we propose. 17 frames match (1) Emotional in FrameNet. They are direct children of the node EMOTIONS (or children of its children), and can thus be considered to have an emotional status in FrameNet. In this, our PMI-based analysis aligns with the database, since it corroborates the intrinsic emotionality of its affective vocabulary. Next, we find that 81 frames, annotated as (2) Stimulus Events, express emotion-inducing circumstances, and 84 frames formalize some components of emotions described by appraisal theories, in line with (3) Appraisal-based Frames.

Compared to (1) Emotional in FrameNet, (2) Stimulus Events and

(3) Appraisal-based Frames, neutral frames tend to express features of objects (e.g., ESTIMATED_VALUE, SUBSTANCE_BY_PHASE, MEASURABLE_ATTRIBUTES) or of events which can be verified and measured objectively (e.g., CHANGE_RESISTANCE, BECOMING_DRY). With respect to the emotional ones, they depend much less on people's own involvement in the state of affairs mentioned in the texts. Some examples are in Figure 6.4 (b).

We now discuss Table 6.4 in more detail.

(1) Emotional in FrameNet. The first group encompasses frames that derive from EMOTIONS. Example children are JUDGMENT, EMOTION_DIRECTED and STIMULUS_FOCUS, FEELING and CONTRITION which express the internal state caused by an emotion episode.

(2) Stimulus Events. Recall that in the view of appraisal theories, events are emotion causes – they make emotions different from other affective states, such as mood, which are more independent from the environment. Our second group of frames captures precisely this notion. It comprises items that revolve around emotion-stimulating circumstances, like ROTTING and DESTROYING, and therefore, can account for the emotionality assigned to texts that convey an affective content via purely factual descriptions. In this light, this cluster is also close to the idea underlying emotion lexicons, namely, that some words evoke mental representations with a prototypical affective substrate, somewhat established in the collective knowledge.

For some of them, an emotional attachment might result weak at first glance, but it is clarified by looking at the texts in which they appear. MAKE_COMPROMISE, for instance, is typically evoked by sentences that bring up people sacrificing self principles; MANIPULATE_INTO_DOING is ascribed to descriptions of bullying episodes; CAUSE_TO_FRAGMENT is evoked by texts depicting an entity being "broken" (e.g., being hurt by a breakup).

Among them are also a few instances that do not indicate events strictly speaking, but kin concepts. VIOLENCE and HOSPITALITY are two examples (associated by the SRL tool to sentences that manifest appreciation for conviviality), as well as IRREGULAR_COMBATANTS, which has to do with fighters and hence with a notion of brutality (comparable to KILLING and BEARING_ARMS). Similarly, MEDICAL_SPECIALTIES is evoked by (potentially stirring) circumstances that are related to healthcare and therapy, and RITE appears in the context of intimate meditations and expressed hopes.

(1) Emotional in FrameNet

Definition: These frames are direct children of the node EMOTIONS. They must be its immediate derivation, or a derivation of one of its children nodes.

Frames: 2. EMOTIONS_BY_STIMULUS, 4. JUDGMENT_DIRECT_ADDRESS, 5. JUST_FOUND_OUT, 6. FEAR, 9. EMOTION_ACTIVE, 12. EXPERIENCER_OBJ, 14. EXPERIENCER_FOCUS, 16. CONTRITION, 20. STIMULUS_FOCUS, 23. MENTAL_STIMULUS_STIMULUS_FOCUS, 29. EMOTION_DIRECTED, 33. JUDGMENT, 35. FEELING, 37. DESIRABILITY, 94. PREDICAMENT, 96. DESIRING

(2) Stimulus Events

Definition: These frames express circumstances that typically cause an emotion.

Frames: 1. DYING, 8. REASSURING, 11. CAUSE_EMOTION, 17. SENTENCING, 18. CAUSE_TO_START, 22. BUNGLING, 26. PROTEST, 32. ROTTING, 40. KILLING, 42. BEAT_OPPONENT, 43. FIRING, 44. DESTROYING, 46. TERRORISM, 47. DARING, 48. VERDICT, 49. FINISH_COMPETITION, 51. OFFENSES, 55. INSTITUTIONALIZATION, 56. DEATH, 57. RECOVERY, 58. SUASION, 61. KIDNAPPING, 62. GUILT_OR_INNOCENCE, 63. CAUSE_TO_EXPERIENCE, 64. SUBJECTIVE_INFLUENCE, 67. CAUSE_HARM, 70. CATASTROPHE, 71. MISDEED, 72. ARREST, 73. PREVENT_OR_ALLOW_POSSESSION, 76. IMPRISONMENT, 78. ARRAIGNMENT, 82. ACCOMPLISHMENT, 83. VIOLENCE, 86. RENDER_NONFUNCTIONAL, 88. FLEEING, 89. UNEMPLOYMENT_RATE, 90. WARNING, 95. ASSISTANCE, 97. IMPROVEMENT_OR_DECLINE, 101. RITE, 102. ENTERING_OF_PLEA, 103. REBELLION, 108. ATTACK, 109. REPEL, 110. HOSTILE_ENCOUNTER, 113. CAUSE_TO_FRAGMENT, 114. DOMINATE_COMPETITOR, 116. RESCUING, 119. PREVARICATION, 122. SUBVERSION, 125. RESOLVE_PROBLEM, 126. EXPERIENCE_BODILY_HARM, 128. ARSON, 130. MANIPULATE_INTO_DOING, 132. MILITARY_OPERATION, 133. MEDICAL_SPECIALTIES, 135. MEDICAL_CONDITIONS, 140. EXAMINATION, 144. INFECTING, 149. RUN_RISK, 154. DEAD_OR_ALIVE, 155. IRREGULAR_COMBATANTS, 158. ENDEAVOR_FAILURE, 159. INVADING, 160. BEING_OPERATIONAL, 161. THEFT, 166. HOSPITALITY, 167. QUARRELING, 170. MEDICAL_INTERVENTION, 171. BEARING_ARMS, 174. REVEAL_SECRET, 178. ESCAPING, 181. DAMAGING, 182. PRISON, 183. MAKE_COMPROMISE, 186. TRIAL, 187. COMMITTING_CRIME, 189. SURVIVING, 192. SURRENDERING, 195. EXECUTION, 200. CURE, 204. ABANDONMENT

(3) Appraisal-based Frames

Definition: Frames capturing the link between emotions and events, their saliency for the well-being of the experiencer, evaluations, actions, motives and responses that the experiencer takes in reaction to emotional events.

3. DISGRACEFUL_SITUATION, 7. MAKING_FACES, 10. FACIAL_EXPRESSION, 13. REWARDS_AND_PUNISHMENTS, 15. FAIRNESS_EVALUATION, 19. COMMUNICATION_NOISE, 21. ACCURACY, 24. LUCK, 25. MENTAL_PROPERTY, 27. MAKE_NOISE, 28. BODY_MARK, 30. SATISFYING, 31. COGITATION, 34. SUCCESS_OR_FAILURE, 36. CHEMICAL-SENSE_DESCRIPTION, 38. FRUGALITY, 39. AGREE_OR_REFUSE_TO_ACT, 41. AESTHETICS, 45. CHAOS, 50. SOCIABILITY, 52. DESERVING, 53. RESPOND_TO_PROPOSAL, 54. CERTAINTY, 59. OMEN, 60. RISKY_SITUATION, 65. BEING_QUESTIONABLE, 66. PROMINENCE, 68. REVENGE, 69. VOCALIZATIONS, 74. BIOLOGICAL_URGE, 75. GRASP, 77. DIFFICULTY, 79. MORALITY_EVALUATION, 80. COMING_TO_BELIEVE, 81. STINGINESS, 84. SOCIAL_INTERACTION_EVALUATION, 85. SUCCESSFUL_ACTION, 87. ARTIFICIALITY, 91. FORGING, 92. RENUNCIATION, 93. HIT_OR_MISS, 98. WEALTHINESS, 99. CORRECTNESS, 100. COMMITMENT, 104. LEVEL_OF_FORCE_EXERTION, 106. COMPLAINING, 107. REASONING, 111. PEOPLE_BY_MORALITY, 112. ENDANGERING, 115. SOCIAL_DESIRABILITY, 118. JUSTIFYING, 120. JUDGMENT_COMMUNICATION, 121. WILLINGNESS, 123. SENSATION, 124. COMPATIBILITY, 127. INCLINATION, 129. EXPRESSING_PUBLICLY, 136. TRIGGERING, 141. EXPECTATION, 142. EXPEND_RESOURCE, 143. JUDGMENT_OF_INTENSITY, 148. TRUST, 151. EVENT, 152. OPPORTUNITY, 153. BEING_RELEVANT, 156. AWARENESS_STATUS, 157. DYNAMISM, 165. FAME, 168. BEING_AT_RISK, 169. OPINION, 172. REQUIRED_EVENT, 177. EDUCATION_TEACHING, 179. CAUSE_IMPACT, 184. PRECARIOUSNESS, 185. MEET_SPECIFICATIONS, 188. MOTION_NOISE, 190. ATTEMPT, 194. BREATHING, 196. CONFRONTING_PROBLEM, 197. EVENTIVE_AFFECTING, 199. ATTITUDE_DESCRIPTION, 201. EXPERTISE, 203. AWARENESS

Idiosyncratic

Definition: Frames that do not belong to any of the above groups.

Frames: 105. LINGUISTIC_MEANING, 117. BOARD_VEHICLE, 131. LAUNCH_PROCESS, 134. REFORMING_A_SYSTEM, 137. ECONOMY, 138. TEMPERATURE, 139. CO-ASSOCIATION, 145. AFFIRM_OR_DENY, 146. BEHIND_THE_SCENES, 147. APPELLATIONS, 150. RIDE_VEHICLE, 162. FUNDING, 163. DURATION_RELATION, 164. CHANGE_OF_LEADERSHIP, 173. PEOPLE_BY_RELIGION, 175. MEDICAL_INTERACTION_SCENARIO, 176. PROLIFERATING_IN_NUMBER, 180. CAUSE_TO_RESUME, 191. MAKE_AGREEMENT_ON_ACTION, 193. REPRESENTATIVE, 198. TOURING, 202. LEGAL_RULINGS

Table 6.4: Partition of emotional frames into groups that capture similar emotion properties. Each frame in numbered with respect to its PMI rank.

(3) Appraisal-based Frames. The third group of frames formalizes implicitly emotional cases like (2) Stimulus Events, but it captures either internal properties of events, as evaluated by the emotion experiencer, or other emotion components that manifest in the experiencer's reaction. Similar to events, also these are given a prominent role by appraisal theories: the emotion mechanism involves an experiencer who assesses the circumstance and engages in a series of changes – i.e., subjective feelings, neurophysiological, motor and motivational alterations.

Frames concerning evaluations are, e.g., SATISFYING and FAIRNESS_EVALUATION. The latter frame, whose link to emotions seemed hazy at first, can in this light be said an emotional exemplar in its own right: the notion of assessment that it brings into play is central to the elicitation of emotions. In this group are also items that qualify events as endangering for the organism (e.g., DIFFICULTY, RISKY_SITUATION), or as fostering its well-being (e.g., LUCK, WEALTHNESS).

Therefore, this class is "Appraisal-based" in a broad sense: the frames it subsumes which refer to an appraisal reveal an evaluation of the quality of a stimulus, but not necessarily one of the 21 dimensions used in Chapter 3. Only some of these frames recall those criteria more closely. One is the coherence of the event with the personal ideals of the experiencer and with societal norms. Frames like FAIRNESS_EVALUATION and MORALITY_EVALUATION convey precisely this type of evaluation. Similarly, GRASP reflects the criterion by which events are appraised in relation to their implications – e.g., Are they relevant to the experiencer's goals? Can their consequences be estimated? It is evoked by textual chunks that involve a cognizer who acquires knowledge about the significance of a given phenomenon and becomes informed to make predictions about it. Events can also be evaluated with respect to the urgency of a reaction (REQUIRED_EVENT), and to the degree to which the experiencers are certain about what is going on (e.g., How well does the experiencer understand what is happening in the emotional situation? (Smith and Ellsworth, 1985)), which is echoed by the frame CERTAINTY.

Focusing on such evaluation criteria, appraisal theories claim that specific assessments of events lead to specific emotion experiences. For instance, a lack of certainty is likely to result in an episode of fear or hope (Smith and Ellsworth, 1985). To an extent, this is accounted for by relations between frames. CERTAINTY, as an example, is inherited by the node TRUST. Therefore, FrameNet relations seem to explain the affective charge of some of these frames that do not stem from EMOTIONS.

We further observe that some frames relate to the effects that emotions have on the organism. BIOLOGICAL_URGE exemplifies the involvement of internal, physiological states that can motivate actions in response to an event. BEING_BROKEN is associated to entities that feel "mulfunctioning" in a metaphoric sense, as a subjective feeling. Other frames correspond to more observable manifestations of the emotion mechanism, such as vocal verbalizations, facial movements, and other diagnostic features that allow people to understand what their interlocutors feel. Frames like MAKING_FACES, FACIAL_EXPRESSION, COMMUNICATION_NOISE (evoked by texts like *"For the first week I cried."*) and MAKE_NOISE seize these components. Yet other frames, for instance REASSURING and COGITATION (a child node of WORRYING), capture external actions or internal attitudes that can occur in the context of emotional situations.

Singular examples that belong to this group are:

- EDUCATION_TEACHING which, somewhat indirectly, points at the idea that an experiencer undergoes a cognitive process, learning new information, similar to COMING_TO_BELIEVE (e.g., *"She was the one who taught me that any situation can be overcome."*);
- REASONING, which often accompanies texts where an evaluation is expressed by means of a dispute described in the text;
- FAME, appearing in sentences with assessments that are either hyperbolic, like *"Believe me it was epic."*, or that concern one's reputation and beliefs, like *"To besmirch her reputation is outrageous"*.

Also the placement of BREATHING, CAUSE IMPACT and LEVEL_OF_FORCE_EXERTION in this group of frames is not self-explanatory. The first two indicate an emotional reaction, (e.g., sighing and slamming a door). The last usually portrays a property of people or events (e.g., feeling fearless and strong, feeling weak). So do also the following frames:

- DYNAMISM, evoked by texts that express the intensity of an experience;
- MEET_SPECIFICATIONS, coupled in text with mentions of personal achievements, or with expressed sensations of fulfilment;

- EXPERTISE, which is interwoven with messages that are either about the familiarity of an entity with a certain topic, or about being skilled at something.

Idiosyncratic Frames. These 22 emotional frames are instances that could be argued to be more affine to the unemotional category. However, the link to emotionality that they display can still be explained by a handful of factors.

Investigating the sentences in which they appear helps uncover some patterns. BOARD_VEHICLE and RIDE_VEHICLE, for instance, are evoked in texts that have to do with embarking on adventures and journeys: these tend to be emotionally-qualified as they often mention personal stances towards such journeys (e.g., if it was pleasant). Instead, LAUNCH_PROCESS, REFORMING_A_SYSTEM, and CAUSE_TO_RESUME characterize texts that express an idea of personal change, of beginning (e.g., *"We may have reformed, but our enemies have not."*, *"I felt revived"*). Put another way, the emotionality of these frames is more linked to their usage in COCA than to their definition. In that, they are similar to some frames in the group (2) Stimulus Events. Still, they do not quite fit in (2) because of the lexical units instantiating them, which are not emblematic of a situation that has an immediate relation to an emotion. Consider, for instance, the difference between CAUSE_TO_START (in (2)), instantiated by "excite", "provoke", "arouse", and RIDE_VEHICLE (that we consider idiosyncratic) evoked by "ride".

Other cases seem to be artefacts of the systematic mistakes of the SRL tool on the emotional sentences. With TEMPERATURE, the automatic role labeler falls short in understanding the metaphoric use of the word "*cool*", for which that frame is usually predicted. In a similar vein, LEGAL_RULINGS is triggered by predicates picturing a judgment, even if the text does not mention any legal context. LINGUISTIC_MEANING is a similar case. It is identified in phrases that are related to meanings and to the "making sense" of a situation, but not in the context of a discussion about linguistic meaning specifically.

Lastly, it is reasonable to think that the positive PMI of some of these frames (e.g., DURATION_RELATION and FUNDING) is inherited from others: in the sentences of COCA, they never appear alone, and therefore, they assume the emotionality of the "company they keep", which is responsible for the texts' emotional content.

Overall, the PMI ranks of these 22 frames show that they are among the lowest values in the top quartile of the distribution – a pointer to their more arguable emotionality.

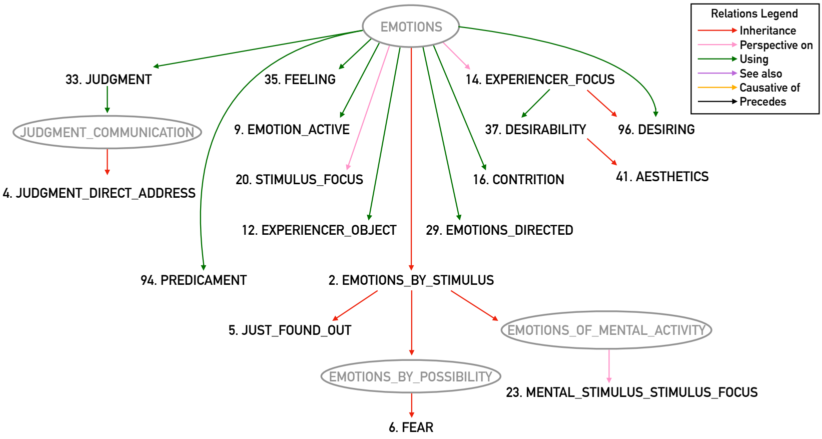


Figure 6.5: Emotional frames (text in black) which are children of the node EMOTIONS, corresponding to (1) in Table 6.4.

We report here a visual prospect of the three emotion groups (excluding the idiosyncratic one). Figure 6.5 illustrates frames that are already encoded as emotional in the database. Figure 6.6 sketches emotion-eliciting events (the frame EVENT is represented there and not in the appraisal group, because it delineates a generic super-category from which all other specific events branch out). Frames that catch emotion properties and the components documented by appraisal theories are displayed in Figure 6.7. These figures show that each set of frames forms a coherent cluster in FrameNet, as an additional validation of our tripartition. For simplicity, we include only the 100 frames with the highest positive emotional associations. Note that the grey nodes are not among the top 100 frames. They are nevertheless illustrated to reproduce the FrameNet structure and account for how the frames under consideration (text in black) relate to one another through relations (represented by the coloured arrows, each corresponding to a specific type of relation). For readability reasons, we do not show relations between all frames.

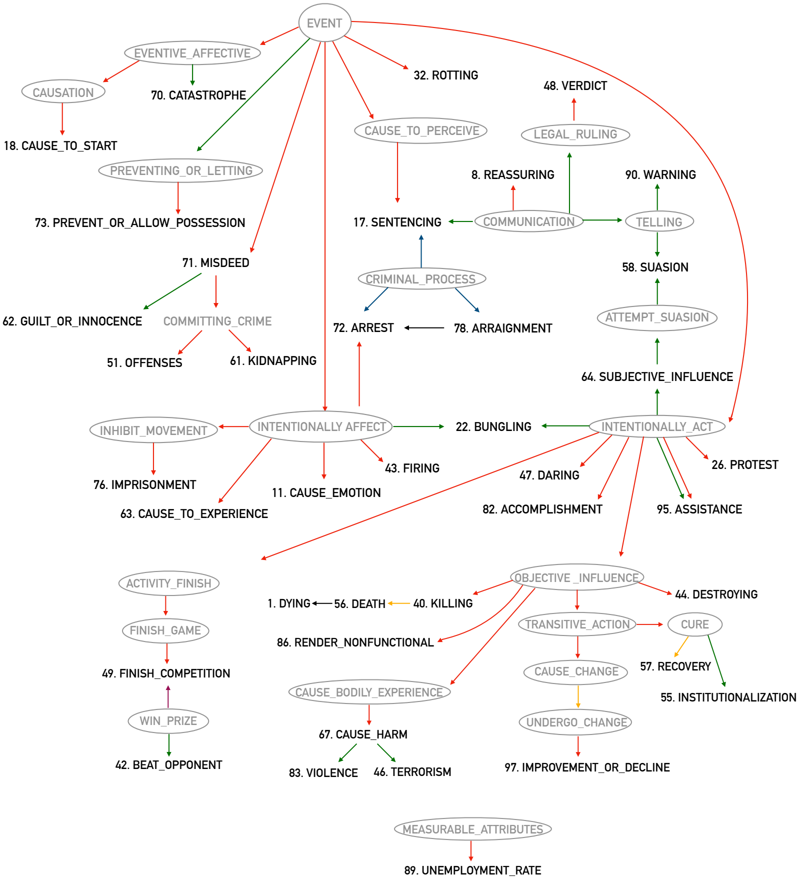


Figure 6.6: Emotional frames (text in black) deriving from the node EVENTS and expressing factual emotion stimuli, extracted from (2) in Table 6.4.

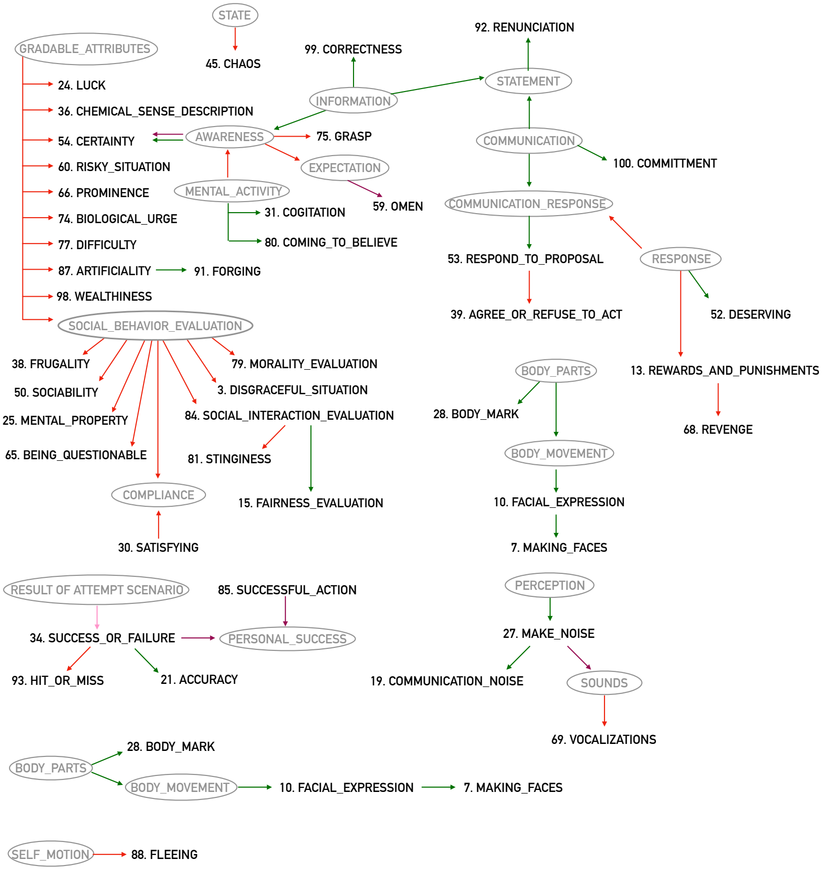


Figure 6.7: Emotional frames (text in black) expressing appraisal-related concepts (cf. (3) Table 6.4).

4.2 Contextually-determined Frames

As already discussed, an emotion association holds for more frames than those explicitly indicated in FrameNet, but it does not encompass the whole database. For one thing, there are neutral frames. In addition, many items have an ambiguous emotional status, as their PMI values fall in the 2^{nd} or 3^{rd} quartiles of the emotional distribution ($-.19 \leq \text{PMI} \leq .23$).

These 408 frames do not present a clear tie with either emotion label. The fact that they can be both emotional and neutral shows an important aspect of the phenomenon under consideration. At times, the relationship that frames hold to their emotion content is underspecified: it is not fixed and bounded to the type of event that they formalize, but rather depends on the overall context in which the frame-evoking predicate appears. Therefore, emotion meanings make no exception in the lexical semantics panorama, where also other phenomena are to be accounted for *in context* (Cruse, 1986) – e.g., word meanings.

With these contextually-determined frames, compositionality is key in the making of an emotion. We see two compositional processes at the sentence level. One is a "within-frame compositionality", in which the predicate is (emotionally) underspecified, but its co-presence with certain arguments can turn out emotional or non-emotional. Illustrative in this regard are sentences like *"I remember this point distinctly."* and *"I remember the magical thinking of my greatest depression."*, both associated to the frame MEMORY but with different predicate arguments (the first sentence is recognized as neutral, the other as emotional). Like in the above examples, many frames are evoked by predicates that serve to introduce topical information, or subordinate sentences. The overall emotionality varies together with the content that they introduce. For instance COMMUNICATION_RESPONSE, TELLING, POINT_OF_DISPUTE, GIVING and GIVE_IMPRESSION have to do with communicative situations that could be loaded with emotionality based on how they are instantiated – what is responded, what is told, what is given (e.g., GIVING in the emotional example *"Cruella gave a gesture of resignation."*). Similarly, UNDERGO_CHANGES pictures some types of transformation, whose outcome could be either emotional or neutral.

The second compositional process that we identify with a manual introspection of the data is an "across-frames compositionality". Frames that appear in combination with a contextually-determined one contribute more to the emotional load of the sentence. The text *"[...] an old girlfriend of mine wrote me this very beautiful letter."*, which is classified as emotional, evokes MEMORY and the emotional AESTHETICS, while *"The words "property value" are ones I remember."*, annotated as emotional by the classifier, evokes MEMORY and POSSESSION. The across-frame compositionality effect can be observed in Figure 6.8, which reports the count of sentences with a contextually-determined frame and one that is emotional (+Frm$_{emo}$), or one that is neutral (+Frm$_{neu}$). From the figure, we see that texts that contain both a contextual frame and one with a positive PMI tend to be emotional; vice versa for the co-presence

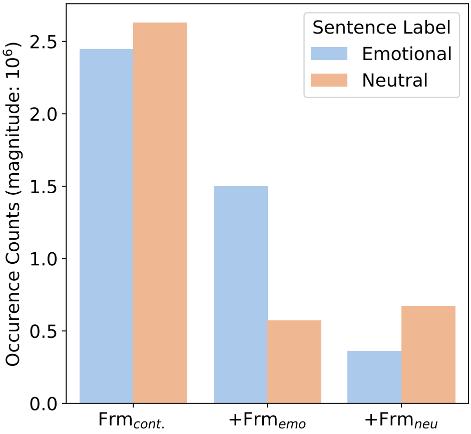


Figure 6.8: Distribution of emotional and neutral sentences evoking contextually-determined frames in isolation (Frm$_{cont.}$) and accompanied by an emotional frame (+Frm$_{emo}$) or a neutral one (+Frm$_{neut}$).

with a neutral frame, found more often in sentences labeled as neutral by the classifier.

Figure 6.8 also shows the distribution of contextually-determined frames that appear alone in a text (Frm$_{cont.}$) across the two emotion labels. They are in >2M emotional and neutral sentences. Devoid of frame interactions, these sentences clarify what it means for frames to be underspecified with respect to emotionality: contextually-determined frames have less to do with properties of things or situations, compared to the neutrally-connotated kins. They rather represent those things (FOOD, VEHICLE, BUILDINGS) or processes (CAUSE_EXPANSION, CAUSE_TO_PERCEIVE). Once more, we notice that when these frames appear in emotional texts, they do so as side information to the main affective meaning, and do not correspond to the predicate that triggers such emotion content. For instance, CONTINUED_STATE_OF_AFFAIRS in the text *"Glad she's still on the show."* is unrelated to the mental state of the subject.

5 Discussion

The PMI analysis disclosed that the relationship between the emotionality of a sentence and that of frames is not straightforward. Frames with a strong positive or negative association to emotionality can be found in texts that overall express the opposite affective content. Even the frames that FrameNet explicitly links to the emotion domain are evoked by neutral sentences. EMOTIONS_BY_STIMULUS, as an example, is found by the SRL tool in the non-emotional *"I had every right to descend this stair, to walk among the glad company [...]"*, because of the lexical unit *"glad"*. Rather than putting the automatic annotation into question, this outcome sheds light on an important fact: sentence-level emotion classifiers can disregard "corners" of emotionality. A verbal expression might have a predominant connotation (e.g., a neutral one, in the example above), which might be correctly identified by the automatic system; yet, by considering entities besides the subject, different emotion nuances emerge (e.g., the company is glad). We already elaborated on this idea in Chapter 3 (Page 101) with a human-assisted analysis. Here, we add that classifiers unaware of the participants affected by an event fail to account for all emotion perspectives, and in such cases the performance of SRL parsers can complement theirs.

Our results bear consequences for our initial hypothesis. We assumed that items in FrameNet can be divided into three groups (neutral, emotional, factual and yet emotional). This turns out to be only partially true because the majority of frames are neither emotional nor neutral. A substantial group is somewhat transparent to emotions: these contextually-determined frames reiterate the need to think about emotionality in terms of relations between words and, as we have seen, of frame compositionality. They raise the question of if and how frames influence the emotionality of a text, and what role they play in a classification task – i.e., To what extent do predicates or arguments contribute to the decisions of an emotion classifier?

Besides contextually-determined frames, the distinction we expected holds. Some frames are clear carriers of emotionality (a finding that matches the current organization of the database), and several frames are emotional despite having a factual denotation. In fact, precisely because they depict concepts that are more descriptive than affective, they pick up on some important components of emotions. They correspond to some of the factors that elicit, underly or manifest an emotion, like events, event evaluations, and emotion effects. The effect components correspond to phenomena that happen in response to emotion-eliciting events (e.g., FACIAL_EXPRESSION) but can also be considered events per se. Consequently, they can evoke specific frames – as hypothesized. Brief, our analysis confirmed that frame semantics captures many levels of an emotion mechanism. In line with previous work (Faruqui et al., 2015, i.a.), it supports the idea that approaches based on embeddings and on human-curated resources cross-fertilize also for emotion analysis.

Our organization of frames in three internally structured groups (plus the idiosyncratic "melting pot") provides a reasonable and theoretically-informed explanation of what different frames have in common from an emotion viewpoint. However, it is an empirical way of grouping them. Even by accepting these three clusters, it should be noted that some items could fit multiple spots. For instance, HIT_OR_MISS and ATTEMPT, which have to do with the goals and concerns of an experiencer (much in an appraisal-oriented fashion), could also be arranged among the emotion stimuli. DESIRABILITY in Figure 6.5, which is dominated by EMOTIONS, expresses a positive stance towards a circumstance and could be placed in the cluster corresponding to Figure 6.7, below GRADABLE_ATTRIBUTES – another node from which DESIRABILITY derives. There is indeed a large number of frames from separate groups that are directly related to one another (e.g., a USING relation holds between MISDEED, which we placed among the emotional stimuli, and MORALITY_EVALUATION).

The proposal of this partition was motivated both by the original structure of FrameNet, and by the fact that both appraisal models and frames make use of a notion of event. Ruppenhofer (2018) had already pointed out that appraisal theories can inform an investigation of the emotion vocabulary in FrameNet. We bolstered that observation by showing the frames to which it extends. Coherent with the insight that appraisals are convenient theoretically (cf. Chapter 2, Section 1.2) and practically for annotation studies (cf. Chapter 3), this chapter showed that they also provide fertile ground to analyze how emotion components reflect in language. However, one could identify other emotion properties and other links to theories different from appraisals, and arrange the emotional frames accordingly.

A logical addition to this study is the use of non-binary emotion labels. We moved to such a setting as a final analysis. This allowed us to carry on the event-oriented approach to emotions initiated in Chapter 3 with annotators, and to link it with frame semantics in a computational framework. Applying the current methodology on crowd-enVENT, we estimated the PMI between frames (identified with the same SRL system used on COCA) and the emotion labels of the descriptions (see Page 86). This emotion-by-emotion analysis has a clear disadvantage:

just like in COCA an event can be emotional or neutral depending on how it is described, the same remark on crowd-enVENT stretches over more than a dozen affective states. It is difficult to interpret why some frames (especially frames that are quite generic, like EXPERIENCER_OBJ), should be conceptually closer to an emotion rather than another. This reinforces the value of having first approached the problem through the (smaller but more focused) lenses of emotionality. Future studies asking if diverse emotion categories traverse FrameNet might look for a way to differentiate the interpretation of these coarse-grained frames, which fit various emotions, from the more fine-grained that by and large resonate with specific emotion names (e.g., WIN_PRIZE: *joy*).

The outcome of the PMI analysis on crowd-enVENT is in Appendix B, Section 4.3. The corpus is too small for the results to be discussed here. More than disclosing the emotions of FrameNet, they summarize what the texts recount, from prototypical situations (e.g., BEAT_OPPONENT, strongly associated to *joy*) to events distinctive of crowd-enVENT (e.g., ANIMALS: *sadness*), and what features they have (e.g., *surprise* is positively associated to JUST_FOUND_OUT, *trust* to SECRECY_STATUS). They exemplify topics and emotion properties noticeable also with a manual analysis of the data – a last signal that our PMI-based method captured meaningful associations between the two variables of interest.

6 Conclusion

This chapter investigated the link between emotionality and meaning. It asked if frames, conceptual abstractions that encode world knowledge, are linked to emotions, and it showed that there is indeed an affective side in the semantics of word relations. More precisely, I described how we automatically annotated COCA with emotion labels and with frames, how we investigated the relationship between them, and how we used PMI to answer our research question.

To conclude, I will point out aspects that differentiate this study from previous approaches, I will sum up its findings, and envision its possible developments.

Difference to Past Work. I moved the focus of typical dictionary-based approaches in computational emotion analysis from individual emotion cues to predicate-argument structures (grasped by SRL tools) as the meeting point between syntax and the U-semantics of Fillmore.

Frames presuppose an acknowledgement of the physical and social world, and they can be evoked by varied lexical units. They therefore capture paradigmatic phenomena and allow to abstract away from the specific terms that instantiate them – and in a way, from style.

By indicating *what* happened and *who* participated in a situation (Fillmore et al., 1976), frames represent the structural components of real-life events that stimulate emotional responses. This aligns the current study with research in structured emotion analysis, which segments texts into semantic roles (Oberländer and Klinger, 2020; Oberländer et al., 2020) and observes emotions with respect to their experiencers and eliciting causes. However, I made use of semantic roles as a means (not as an ultimate structured prediction output) to elucidate the emotion import of texts.

A work that has a similar goal exists: Balahur and Tanev (2016) perform emotion detection based on properties of events, such as actors, patients and action types. While they used a dedicated knowledge base that includes emotion-related information (without any reference to frames), my study aimed at understanding whether such information is already contained in a well-established resource for semantic role labeling.

This analysis set itself apart from work in emotion analysis (e.g., Abdul-Mageed and Ungar, 2017; Felbo et al., 2017; Demszky et al., 2020) also in other respects. It dealt with emotionality instead of a fine-grained set of emotions. Moreover, for the first time in the field, it brought the strengths of frame semantics and appraisal formalisms together.

We could have adopted the OCC model (Ortony et al., 1988) as an alternative event- and evaluation-oriented framework, already proven useful for the study of textual emotions (Shaikh et al., 2009; Udochukwu and He, 2015); but at a deeper look, it does not represent a sufficient theory in this context, because it sees emotions as a descriptive structure of prototypical situations. Its constructs, for instance the binary evaluations, are purely conceptual, and have little to do with linguistic expressions of events. It is thus ill-suited for frame semantics, which is primarily linguistic and does not necessarily match conceptual considerations about the logic-like processes of the OCC model. Appraisal theories that received less attention in computational emotion analysis offered a more viable ground to discuss emotions in language. Specifically, we followed the theoretical system of Scherer (1984), which explains emotions as processes involving the subsystems of an organism (cognitive, motivational, motor, etc.), and which has a theoretical linguistics counterpart (see the theory of appraisal in language, briefly discussed in footnote 9, Page 32).

Findings and Limitations. Results showed that there are frames with a prominent emotional import. Some of them are direct children of the frame EMOTION present in FrameNet. Many others have a less intuitive relation to an affective state, but they still reflect components of emotions spelled out in the psychological literature. This provides quantitative evidence that emotionality is a dimension of meaning that frames possess, even though it is not a piece of information directly provided by the database. In addition, the qualitative analysis emphasized that individual predicates do not always carry the same type of emotion load. On the contrary, their import can depend on the context in which the predicate is situated, namely, on syntagmatic facts.

The study was conducted on a corpus whose size and textual variety is advantageous. It permitted to observe the occurrence of frames across a wide range of emotion-bearing and neutral expressions, and therefore, to generalize our empirical observations. Still, some design choices in the analysis have limitations that we need to acknowledge. Our annotation looked at emotions as a dichotomy emotional vs. neutral, and our PMI-based investigation was not suitable to observe if different frames carry specific emotions. Second, to measure the association with emotionality, we treated all frames equally and as separate entities. While transparent, this choice does not account for within-sentence frames interactions. It further adds noise on the resulting PMI distribution: ignoring frames that carry little emotion meaning could have returned a more robust picture of the association between the two variables. The mapping between these phenomena is indeed not trivial. As we have seen, contextually-determined items, which constitute much part of FrameNet, merely serve as "embedding" frames that introduce (possibly emotional) topics or other frames.

How to Proceed. The salient features of frames that this work revealed open up potential ventures for frame semanticists. Future FrameNet developments could specify what frames carry emotionality with the use of *semantic types*. Semantic types mark general properties of frames and semantic roles, such as variations in the speech use of different lexical units, which could not otherwise be understood from the resource. In FrameNet there already exists a semantic type that is close in spirit to emotions. It indicates the polarity of lexical units like "*compliment*" and "*reprimand*", both of which instantiate JUDGMENT_DIRECT_ADDRESS and whose valence is indicated by the semantic types “Positive_judgment” and “Negative_judgment”. It would be possible to adopt the same idea for the three semantic groups proposed in this paper, or for similar partitions. We refrained from modeling this information into FrameNet ourselves – an endeavor that would require careful and lexicographically motivated annotation, which exceeds the scope of the study.

Future work can also look at contextually-determined frames more in depth. This chapter has identified two compositionality processes that characterize them, but without any robust and empirical substantiation of their functioning. That could involve, for example, repeating our study on different groups of texts: those evoking only one frame, and those containing frame sequences of a predetermined length (e.g., sequences of two frames) that frequently appear in the corpus, with one frame being highly emotional or highly neutral. The objective of a similar setup is to test the phenomenon of across-frame compositionality by observing if contextually-determined frames exhibit consistently higher PMI values when sided by emotional items, compared to when they appear in isolation. Vice versa for their co-presence with the nonemotional counterpart. The results could then be correlated with human-based annotations of frame emotionality in context, as to validate whether the PMI values found in different frame sequences reflect shifts in the emotional load perceived by readers. Moreover, with a qualitative exploration, one could focus on separate contextually-determined frames, and ascertain if among their “emotional companions” there are some that amplify their emotionality more than others – e.g., Do all Stimulus Events contribute to increasing it in the same way? Are certain frame combinations more likely to result in the absence of emotionality?

The insights presented here further inform computational emotion analysis, where research could build systems that are simultaneously emotion- and frame-aware. The frames-to-PMI association scores resulting from this work come handy for that purpose. An interesting track to explore is the contribution of different parts of texts (e.g., frames, arguments, other words) on automatic emotion predictions – e.g., Do classifiers attend predicates to the same extent when judging a text that evokes an emotional frame and a text that evokes a contextually-determined frame? Lastly, research in the field that plans to follow appraisal theories can leverage the *fil rouge* that intersects frame semantics, psychology, and emotion analysis: frames model the verbal expressions of emotion components, among other events, capturing the multiple and nuanced realizations through which embodied emotions and the cognitive evaluations underlying them appear in language. In this regard, we sketched an empirical mapping from frames to a few appraisal dimensions, but it would be important to take the reverse direction as well. Understanding if frames cover all cognitive appraisal dimensions documented by the theories could tell us if frame-based SRL can be applied as an identification strategy of such dimensions, namely, of the criteria that humans use to evaluate events, that lead to an emotion episode, and that also emerge from text. In that case, frames could be used as input to computational emotion analysis pipelines as an effective source of world knowledge information.

Chapter 7

Conclusion

This thesis has been concerned with critical issues in the field of computational emotion analysis, which reveal the need for a more robust understanding of the object to investigate in text:

- First, the need to understand how humans handle implicit emotions as a prerequisite to designing systems that do the same. Overt linguistic markers are often omitted in emotion expressions, and this makes automatic classification difficult.
- Second, the need to clarify how to interpret differences in human judgments, because emotions are subjective and no inference about them is irrefutable.
- Lastly, the need to establish whether emotions lie in the style or in the meaning of texts, as a first step to shedding light on the linguistic level from which humans extract affective information. Determining which of the two perspectives is the most valuable or defensible was relevant also from an applicative standpoint. The recent explosion of the generation task of text style transfer led to an operational emotion–style identity. However, the mutual exclusivity of this identity with the emotion–meaning one was left unverified, and so was the feasibility of the emotion style transfer goal.

I now go through the questions posed in Chapter 1. For each of them, I structure the discussion as follows: I spell out the answers of the dissertation; I summarize my contributions to the field, contextualizing them in the broader scope of NLP; I formulate new questions based on my findings, outlining possible avenues for future research.

1 RQ1: How Well do Humans Recognize Emotions from Implicit Expressions?

1.1 Answer

Unsurprisingly, they are “imperfect” with respect to agreement, but there are appraisal patterns underpinning different emotion labels, which act as a meaningful justification for the coders’ choices. Appraisal ratings reveal in great detail how annotators evaluated the described events, explaining why they inferred a specific emotion, and why not the one that prompted text generation.

1.2 Contributions

Chapter 3 answered the first research question by investigating crowd-enVENT, deISEAR and enISEAR, three resources designed in analogy to the well-established ISEAR dataset. While developing a method that gives insight into the human recognition abilities in regard to implicit expressions, we also resolved the absence of emotion ground truths provided by first-hand emotion experiencers.

We showed the benefit of collecting annotations in a framework that reflects the production and decoding of emotions beyond language. The framework took place around a two-fold proposal. One was the adoption of appraisal variables, a support annotation layer that pushed the field towards other theories than the discrete and dimensional models currently in use. The second pillar of the framework was the employment of a communication model as a backbone for our experimental design. The annotation schema, which comprised writers and readers in a multilingual setup, proved valuable for analyzing the data and observing inter-annotator agreements within the group of readers and between them and the writers. We applied it to label recollections of events, but it can be enlarged to other types of implicit expressions (e.g., emotion-centered metaphors).

1.3 Future Avenues

Indirectly, my first research question urged us to widen our understanding of the type of data that is suitable for classification. Can we model emotions from factual statements? We can, according to our results that all in all spoke in favor of a "rationality" of emotions, revealing the consistency between the answers of multiple people. Still, the way we delimit our object of study needs careful consideration. Below, I outline how to address this point, both theoretically and practically, in the drafting of our guidelines and the creation of our models.

Theoretical Understanding of Textual Emotions. The fact that people perceived emotions even in the corner case of unemotional events has been a notable finding in our study. It is important in two respects. First, it renders evident a difficulty inherent to the study of emotion recognition abilities from text: the trouble to tear apart the contribution of the text itself, i.e., how it presents and qualifies an event, from the prior knowledge that the coders had on such event. Second, it suggests that there is an emotional side to literal meanings (and as we saw, that side is captured by frame semantics) which does not necessarily correspond to the intended one.

Concerning the first aspect, the contribution of both the description and knowledge of an event in recognizing emotions, it seems clear that the too often purported statement in the field that *texts convey emotions* is a simplification, at best. To define our object of research better, it might be useful to make the following distinction. On the one hand, linguistic realizations "permeated" by an emotion, as it were, which contain the sufficient and necessary information to make an emotion inference. On the other, those for which the judges *assume that the text carries an emotion implicature*, and pick a label based on that assumption. In such cases, their annotations disclose their mental representations of events more than the emotion load of language.

For the second aspect, a question to ask is whether implicit expressions produced under the *will to transmit* an emotion differ from those that were analyzed in this dissertation. Its answer also carries implications for data collection studies, as outlined next.

Data Collection and Annotation. Information gathered from the text authors allowed us to bypass a problem on which emotion analysis stumbles. The practice of tasking coders to infer the writers' emotions (without having the answers of these) seems to presuppose that writers produced their verbal signal while feeling that emotion, or even, while being pushed to communication by it. We have seen that this is not the case. The validators in our experiments reconstructed the emotions of people whose mental state did not necessarily correspond to the one prompting event recollection.

That is an important criticism because it casts doubt on whether the phenomenon we study in computational emotion analysis is the one we define in our guidelines. Moreover, it raises the question of how different the readers' recognition skills would have proven, had my studies examined texts written with a spontaneous drive to transmit an emotion. To study that, a different setup than ours (based on recollections of past states) is required, and more attention should go into pragmatics, a whole other dimension of language that I have dismissed, which is where the intentions behind the verbalization of emotions instantiate to, e.g., sway, persuade, affect, or simply inform our interlocutors.

The desirability to regard communicative variables other than linguistic signals also echoes another of our findings. The mapping from a text to an emotion is not unique but is parametrized by multiple factors. Some of them are linguistics, others are prior to language (we called them in Θ_i). Chapter 4 focused on the validators and observed how their emotion classification performance changed at varying Θ_i conditions. An alternative research route would be to verify if the readers' decisions change if they knew who wrote a given text. The emotion encoding specific to a writer could be better decoded by readers informed about her identity (i.e., the text generator's Θ_i). By focusing instead on text generation, one could examine if the writers' productions change depending on their expected audience (i.e., the text validators' Θ_i).

Automatic Modeling. One reason that makes the writers' judgments an essential bit of information is that emotion classifiers are part of the ambitious program of affective computing. To have systems interact with humans, the second-hand emotion labels of readers are a suboptimal source of training, since external judges are not always correct. The broader picture would rather benefit from learning within-writer patterns, which would allow to tune a system's decoding ability on the unique way each user expresses her affections.

At the moment, such a person-specific scenario is not a parsimonious solution at each researcher's fingertip, and appropriate datasets for this goal are unavailable. Still, a feasible strategy to improve the models exists, and that is to tie them more strongly to emotion theories in psychology. Like works that integrate dimensional models of affect and discrete categories, we joined appraisal ratings with discrete emotion labels, seeing the dimensions of the former as a latent representation of the latter. Compared to a dimensional model of affect, they have a clearer mapping to emotions; similar to it, they represent structured differences among emotions. Appraisals further surpass the expressiveness of the OCC model, which is applied in emotion classification to make deterministic decisions, disregarding probability distributions across different event evaluation criteria. Under the assumption that appraisals *are* emotions (cf. Chapter 2), modeling can even take place without deciding on the possible set of emotion outputs.

Appraisals, mostly dismissed in the literature of affective computing in text, formalize a primary emotion resource. As such, they can make the systems more human-like and theoretically grounded, possibly achieving better classification performances.

2 RQ2: How Do In-text and Beyond-text Factors Affect Human Emotion Recognition?

2.1 Answer

To the degree to which such factors are variables quantified by the annotators themselves, they cause systematic labeling behaviors that correlate with a change in the annotators' emotion recognition performance. They prompt higher or lower agreements, noticeable both among readers, and between them and writers.

2.2 Contributions

We have discussed the subjectivity of the readers' judgments at length, demonstrating that it is possible to understand disagreements, rather than treating them as an insurmountable flaw of emotion annotations.

In Chapter 3, we have compared the emotions prompting the writers with the readers' through the appraisals rated by both parts. Further, we have looked at how the readers' agreement on appraisals increases as they are informed about the emotion originally elicited by an event. Focusing on the relationship between emotions and appraisals binds the investigation of agreement to a predefined number of event evaluation criteria. Therefore, in Chapter 4, we moved on to formalizing the idea that the mapping from text to emotion is parametrized by Θ_i, which encompasses factors that describe the identity of an annotator, and which accounts for the non-independence of such a mapping from the person carrying it out.

We have selected potential covariates of disagreements, i.e., stable features of the coders, contextual factors and text-related factors, bridging previous NLP studies on annotations with psychological research on emotion recognition. We have successfully identified the impact of these factors on the annotators' appraisal and emotion recognition performance: demographics, personality traits, age, gender, event familiarity, current emotion state, education and ethnicity turned out to be sources of significant agreement differences, both within and between property-specific groups. *Having* particular properties and *sharing* particular properties are conditions that spur more consistent answers.

It would be tempting to generalize our findings and conclude that these factors influence the task of emotion recognition, but that would be incautious. While their effect was visible on our dataset, we did not work with a comparable (and comparably high) number of people for all properties. We had insufficient evidence to understand *how* some factors have an impact on agreement (e.g., Does a state of high anger improve or worsen its score?), nor did we investigate why. Hence, we treat our outcome as a departure point to study each property better in the future, ideally with diverse types of texts, not confined to event descriptions, a balanced number of annotators per trait, and larger-scale datasets.

Narrowing the selection of factors down to a few that pertain to the annotation task, we conducted a systematic evaluation of confidence and intensity, both capable to capture variations in agreement. Lower intensity and lower confidence correlate to a weaker agreement. Importantly, our experiments have demonstrated that people predict when their labels will diverge. This is a promising insight, which manifests that some disagreements are systematic and explainable, and do not necessarily undermine the reliability of annotators.

2.3 Future Avenues

There is ongoing interest in shifting the dominant NLP research paradigm aimed at resolving disagreements towards a perspectivist framework. The idea is to include the opinions of annotators in a more comprehensive manner, in the spirit of accounting for their diverse annotation voices (Friedman et al., 2006). To do that, it is possible to opt for non-aggregated data collection methods (Checco et al., 2017; Plank et al., 2014b; Aroyo and Welty, 2015), recently applied also in the domain of emotions (Ngo et al., 2022). In this context, two relevant questions come up. One is how my findings can effectively support other studies that pursue resource curation. The other is what they teach us about data modeling. Indeed, RQ2 encourages us to think about how to better evaluate the decisions made by automatic classifiers. Just like humans', labels that are erroneous and lower the systems' performance might be qualitatively acceptable.

Data Annotation. The subjectivity of emotions challenges the current notion of reliability, as one that signals a comparable comprehension of the guidelines among coders. In fact, it puts into a critical position even the necessity to look at agreement measures in emotion analysis. Conceding that they are still a useful tool in the field, it is important to notice how some disagreements persist under certain conditions. The readers' choices could contradict one another, but at least part of them evades the umbrella explanation of "plain annotation noise". To differentiate the two cases (i.e., systematic labeling patterns vs. errors), annotation endeavors can collect disagreement covariates, which will help understand what annotation decisions are to be included in the final dataset usable for machine learning purposes, and what is worth discarding. In general, the major message of our findings is that broadening the annotation scope is a good idea, because it reflects the complexity of phenomena that involve linguistic and extralinguistic knowledge. Among all variables we considered, uncertainty is portable to other emotion annotation endeavors as well as other phenomena besides affect.

Automatic Modeling. Taken together, my observations on disagreement put into sharper focus the ultimate goal of creating systems that display an emotional intelligence similar to humans (Picard, 2000). Needless to say, we are far from automatizing many aspects of our emotion recognition skills. Systems do not know *how* anger, joy, hope *feel like*. They lack the necessary network of cognitive, social and motor abilities involved in an emotion episode. Most importantly, they lack the mind–body feedback, which allows humans to persist in a mental state with an embodied experience of it (Minsky, 2006). That is one reason for which we understand what other humans feel, and the reason why researchers can refrain from defining emotions in the annotation guidelines, trusting that the coders will tap on their knowledge about those conditions.

This limitation does not undermine machines in their possibility to achieve a certain degree of empirical proficiency in the text–to–emotion mapping task. Hence, what does it means for an emotion classification model to be a good model, in light of the fact that even human readers are flawed (i.e., discordant with one another and with the ground truth of the emoter)?

My proposal is to look beyond emotion labeling performances: a good model is one that *justifies its choices* like humans. Other than appraisals, a justification can be given by the behavior that the systems expect from the utterer of a text. Expressions of emotions are signs of a class of mental conditions that make our behavior more predictable (Minsky, 2006). Investigating if observers (human or otherwise) are able to estimate what the emoter would do (or say) next is a promising way to evaluate whether they inferred the correct emoter's state. Any eventual alignment between an emotion decision and the predicted behavior (e.g., ⟨ *anger*, attack ⟩) could reveal the reasonableness of the classification output, even when the latter does not match the emoter's; and in case this ideal classifier predicted the correct behavior while picking the wrong emotion label, that would still account for a phenomenon that occurs among humans as well, namely, successfully sensing the consequences of what others feel but wrongly naming the sensation.

Throughout the thesis, I have sought to close the gap between computational analysis and psychology. The approach just sketched respects this condition by including a notion of behavior. The motivational tradition emphasizes indeed that emotions *serve* a function, e.g., to intimidate, to get closer, etc. They move us. Therefore, if there exists a relationship between successful communication and emotion decoding, it can be measured beyond words: how much this emotion information I am inferring changes my internal state, my beliefs, my ability to predict what the message sender will do, her actions, what they imply for my well-being and utility.

3 RQ3: Where Are Emotions?

3.1 Answer

In the meaning of sentences, not in their style – but only to the extent that the two dimensions can be treated as independent aspects of language.

3.2 Contributions

We answered this question using style transfer and frame semantics as tools to investigate style and meaning, respectively.

Chapter 5 addressed the style transfer task via backtranslation in a progression of experiments. Finding that neural MT loses emotion information from one language to another and back, we proposed to re-rank the translation hypotheses with the help of an emotion classifier. The procedure fixed the problem of emotion loss and was employed to promote emotion variability for style transfer. Compared to architectures trained end-to-end for the purpose of style transfer (e.g., Smith et al., 2019), our approach is more interpretable (the re-ranking gives plain transparency over why an output is preferred over others), easier to implement (it only requires establishing the re-ranking function, while fluency and naturalness are maximized by the translation system), and more versatile (it can be leveraged for inducing both emotion recovery and emotion variability). It also holds up with the ability of those methods, whose re-styled paraphrases tend to boil down to emotion names substitutions (input: *"I am happy I did . . . "*, output: *"I am angry I did . . . "*), hardly maintaining input meanings. The qualitative evidence we obtained rules out the full attainability of this goal (i.e., to alter the emotion attribute of a text without disrupting the input semantics), and the idea that emotions reside in the style of sentences.

Chapter 6 moved on to investigating meanings. It started from the remark that the interpretation of words can be guided by extralinguistic knowledge. This motivated us to take distance from the traditional, lexical-based approaches of emotion analysis, and have a better look at the role that background information plays in emotions understanding, via frame semantics. While a small part of FrameNet is ostentatiously linked to emotions (e.g., FEAR), we did not make any assumption as to which frames are emotional, but we exploited an automatic procedure to identify them systematically. By investigating the mutual information between the emotion categories of sentences and the frames that they evoke, we provided evidence that emotionality is an integral part of the meaning of frames (by association, not by definition) besides those made available in FrameNet as emotion-related ones. We further showed how emotional frames reflect concepts inherent to the definition of emotion in the theoretical framework of emotion appraisal from psychology. A dictionary of frames-to-emotion associations is the output of this analysis with an applicative value. It can be employed in alternative to typical word-to-emotion lexicons.

3.3 Future Avenues

We approached the third research question with computational tools to operate on a larger scale than allowed by human-assisted experiments. Our findings represent a compass to start understanding the level of a linguistic signal where emotions are looked for by people, when they employ their recognition abilities studied for RQ1 and RQ2. The need for human annotators can mark an onset for future work, as do some observations drawn at different points of the thesis. In what follows, I establish immediate steps to take, while connecting style transfer and semantic role labeling seen for RQ3 with ideas derived from other chapters.

Style Transfer. A promising hypothesis to explore is that emotion style transfer is to be conducted with a different definition of "style" or with different style transfer desiderata. Potential directions to do so are:

- Addressing style transfer through emotion perspectives: Chapter 3 showed that the event described in a text elicits specific emotions and appraisals by putting oneself in the shoes of different entities involved in the event. The goal of style transfer could be phrased as one of emotional storytelling that varies the narrative perspective from one entity to another. The desideratum of preserving meaning leaves space to that of preserving the input event; the objective of attribute transfer consists in putting the aspects of an event that convey the target emotion (i.e., the one salient for the target narrator) in the foreground.
- Addressing style transfer with Θ: factors that Chapter 4 has found to influence emotion recognition are the factors that the style transfer literature sees as underlying the style of texts. It

is therefore reasonable to think they shape how emotions are encoded in language. Style transfer could target them all, simultaneously, and observe whether they have a stronger impact on the emotion of the resulting paraphrases than when the transfer of such emotion is targeted in isolation.

- Addressing style transfer with frames: once the relationship between frames and fine-grained emotions (not emotionality, as in Chapter 6) will be established, the style transfer paraphrasing goal could be integrated with the ability of frames to grasp structural aspects of meaning. That is, frames provide a latent representation that has to remain invariant between input and output, while being instantiated in the two parts with differently-connotated lexical units that evoke the same frame.

Emotions might not be stylistic attributes to transfer in the succinct verbal productions I have dealt with, but the task could be reconsidered by upscaling the linguistic context in which texts occur, and where the writers' unique style can more easily emerge. With the amount of evidence we have, I believe that a complete rejection of the relationship between emotions and style would be a faux pas. For *what* elicited an emotional reaction in us one time (i.e., the denotation of a described event or event property) can leave us neutral another time. Our affective stance, our contextual judgments of value come to light from *how* we present it.

Semantic Role Labeling. To close the appraisal circle opened in Chapter 3 with the ability of frames to capture emotion properties (emerged in Chapter 6), upcoming emotion SRL activities can incorporate appraisal information. Emotions are part of a large slice of FrameNet. Therefore, informing the systems about the potential evaluations conducted by the event participants will plausibly improve the automatic understanding of what happened, and what properties it had.

4 Final Remarks

In conclusion, this thesis has employed computational linguistics methods that involved humans and automatic systems generating and classifying text. It has lifted the NLP perspective, predominant in all studies, with the support of theories from psychology and linguistics. Their tools served to test and strengthen the foundations of the edifice of computational emotion analysis. My main goal was to improve our comprehension of emotions in language. This comprehension now pays psychology back. It corroborates that the findings that inspired our design decisions, as well as the notions that we borrowed from scientists focused on non-verbal stimuli, generalize to text, when looking at emotions from the computational angle of NLP.

Should the end ambition of building an emotion-aware machine retain one message from this dissertation, it would be to refrain from proceeding in separate compartments. To study and automatize a phenomenon that emerges in and outside of language, a trans-disciplinary approach can be nothing but beneficial.

Appendix

Appendix A

Guidelines and Data Collection Details

1 Emotion Detection from Implicit Expressions

For experimental reproducibility, I detail here our approach to building crowd-enVENT, enISEAR and deISEAR. Differences between the annotation variables in crowd-enVENT, enISEAR and deISEAR are in Section 1.1. For crowd-enVENT, I compare our questionnaire to the original studies in psychology on which it is based (Section 1.2), and I provide details about crowdsourcing (Section 1.3), and I report our guidelines for the generation and the validation phases (Section 1.4). Details on the experimental setup of enISEAR and deISEAR are in Section 1.5, the guidelines used to perform the manual analysis of events are in Section 1.6.

1.1 Comparison of Guidelines for enISEAR and crowd-enVENT

Phase 1 was conducted with a similar approach in (de/)enISEAR and crowd-enVENT. However, the variables used for the first were a subset of the other. Table A.1 visualizes them side by side.

Variables that appeared in both studies differed in the answer options. In en(/de)ISEAR: prompting emotions were Ekman's plus *shame*;

	In de(/en)ISEAR	In crowd-enVENT	Different answer options
Prompting Emotions	✓	✓	✓
Current Emotion		✓	
Appraisals		✓	
Temporal Distance	✓		
Emotion Duration	✓	✓	✓
Intensity	✓	✓	✓
Event Duration		✓	
Confidence		✓	
Gender	✓	✓	✓
Age		✓	
Education		✓	
Ethnicity		✓	
Personality Traits		✓	
Event Familiarity		✓	

Table A.1: Variables used for de(/en)ISEAR and crowd-enVENT.

emotion duration was one among "a few minutes", "an hour", "several hours", "a day or more"; *intensity* is rated on a 4-point scale ("not very", "moderately intense", "intense", "very intense"); *gender* could take on 3 values ("female", "male", "other"). In crowd-enVENT: prompting emotions were the 12 emotions selected in Chapter 3, Section 3.3.2 plus the *noemotion* label; *emotion duration* was rated on a 5-point scale, and so is *intensity*; *gender* could take on 4 values ("female", "male", "gender variant/non-conforming", "prefer not to answer"). The full list of possible answers for crowd-enVENT is detailed in the example questionnaire below (Section 1.4 of this Appendix).

1.2 crowd-enVENT: Comparison of our Appraisal Dimensions Formulations to the Literature

Table A.2 reports a comparison of the appraisal statements used in the generation phase of crowd-enVENT with the original formulations in Scherer and Wallbott (1997) and Smith and Ellsworth (1985). The statements were rated from 1 to 5 (with 1 being "not at all" and 5 "extremely"). Similarly, answers for Scherer and Wallbott (1997) were picked on a 5-point Likert scale between "not at all" over "moderately" to "extremely", with an addition option "N/A". Smith and Ellsworth (1985) chose a 11-point scale.

Dim.	SW/SE	crowd-enVENT
Relevance Detection: Novelty Check		
Suddenness	SW: At the time of experiencing the emotion, did you think that the event happened very suddenly and abruptly?	The event was sudden or abrupt.
Familiarity	SW: At the time of experiencing the emotion, did you think that you were familiar with this type of event?	The event was familiar.
Event Predictability	SW: At the time of experiencing the emotion, did you think that you could have predicted the occurrence of the event?	I could have predicted the occurrence of the event.
Attention, Attention Removal	SE: Think about what was causing you to feel happy in this situation. When you were feeling happy, to what extent did you try to devote your attention to this thing, or divert your attention from it.	I paid attention to the situation. I tried to shut the situation out of my mind.
Relevance Detection: Intrinsic Pleasantness		
Unpleasantn. Pleasantness	SW: How would you evaluate this type of event in general, independent of your specific needs and desires in the situation you reported above? Pleasantness Unpleasentness	The event was pleasant for me. The event was unpleasant for me.
Relevance Detection: Goal Relevance		
Relevance	SW: At the time of experiencing the emotion, did you think that the event would have very important consequences for you?	I expected the event to have important consequences for me.
Implication Assessment: Causality: agent		

...continued

Dim.	SW/SE	crowd-enVENT
Own, Others', Situational Responsibility	SW: At the time of the event, to what extent did you think that one or more of the following factors caused the event? Your own behavior. The behavior of one or more other person(s) Chance, special circumstances, or natural forces.	The event was caused by the my own behavior. The event was caused by somebody else's behavior. The event was caused by chance, special circumstances , or natural forces.
Implication Assessment: Goal Conduciveness		
Goal Support	SW: At the time of experiencing the emotion, did you think that real or potential consequences of the event... ... did or would bring about positive, desirable outcomes for you (e.g., helping you to reach a goal, giving pleasure, or terminating an unpleasant situation)? ...did or would bring about negative, undesirable outcomes for you (e.g., preventing you from reaching a goal or satisfying a need, resulting in bodily harm, or producing unpleasant feelings)?	At that time I felt that the event had positive consequences for me.
Implication Assessment: Outcome Probability		
Consequence Anticipation	SW: At the time of experiencing the emotion, did you think that the real or potential consequences of the event had already been felt by you or were completely predictable?	At that time I anticipated the consequences of the event.
Implication Assessment: Urgency		

...continued

Dim.	SW/SE	crowd-enVENT
Response urgency	SW: After you had a good idea of what the probable consequences of the event would be, did you think that it was urgent to act immediately?	The event required an immediate response.
Coping Potential: Control		
Own, Others', Chance Control	SE: When you were feeling happy, to what extent did you feel that you had the ability to influence what was happening in this situation? someone other than yourself was controlling what was happening in this situation? circumstances beyond anyone's control were controlling what was happening in this situation?	I had the capacity to affect what was going on during the event. Someone or something other than me was influencing what was going on during the situation. The situation was the result of outside influences of which nobody had control.
Coping Potential: Adjustment Check		
Anticipated Acceptance	SW: After you had a good idea of what the probable consequences of the event would be, did you think that you could live with, and adjust to, the consequences of the event that could not be avoided or modified?	I anticipated that I could live with the unavoidable consequences of the event.
Effort	SE: When you were feeling happy, how much effort (mental or physical) did you feel this situation required you to expend?	The situation required me a great deal of energy to deal with it.
Normative Significance: Control		
Internal Standards Compatibility	SW: At the time of experiencing the emotion, did you think that the actions that produced the event were morally and ethically acceptable?	The event clashed with my standards and ideals.

…continued

Dim.	SW/SE	crowd-enVENT
External Norms Compatibility	SW: At the time of experiencing the emotion, did you think that the actions that produced the event violated laws or social norms?	The event violated laws or socially accepted norms.

Table A.2: Comparison of formulations of items between Scherer and Wallbott (1997) (SW), Smith and Ellsworth (1985) (SE), and crowd-enVENT.

1.3 crowd-enVENT: Corpus Generation and Labeling

The creation of crowd-enVENT took place over a period of 8 months (from March to December 2021). We organized data generation into 9 consecutive rounds. A round was aimed at collecting a certain number of tasks, based on different emotions. The first round served to verify whether our variables were understandable, record the feedback of the annotators, and adjust the questionnaire accordingly. We do not include it in crowd-enVENT. The three final rounds balanced out the data. They comprised questionnaires only for the emotions with insufficient data points, due to rejections in the previous rounds.

We enlisted participants via Prolific, a platform that allows to prescreen workers based on several features (e.g., language, nationality). We opened the study only to contributors with a nationality from the US, UK, Australia, New Zealand, Canada, or Ireland. To boost the quality of crowd-enVENT, we opened the crowdsourcing study only to participants with an acceptance rate of ≥80% to previous Prolific jobs, for both phases. We further interspersed our questionnaires with two types of attention tests: a strict test, in which a specified box on a scale had to be selected, and one in which a given word had to be typed. We aimed at making automatic text corrections unlikely, by impeding the completion of our surveys via smartphones.

For the generation phase, workers were provided with a list of generic life areas (i.e., health, career, finances, community, fun/leisure, sports, arts, personal relationships, travel, education, shopping, learning, food, nature, hobbies, work) that could help them pick an event from their past, in case they found the task of choosing one troublesome. These answers were collected on Google Forms. Annotators could fill in more than one questionnaire (for more than one emotion, in more than one round). On average, people took our study 2.8 times, with the most productive worker contributing with 33 questionnaires. Since our expected completion time for a questionnaire was around 4 minutes, we set the payment to £ 0.50, i.e., £ 7.50 per hour, in the respect of the minimum Prolific wage.

We discarded submissions with heavily ungrammatical descriptions and incorrect test checks (i.e., those based on box ticks), while we were lenient with type-in checks containing misspellings. For individual annotators who completed various questionnaires, we removed descriptions paraphrasing the same event. The 6600 approved questionnaires were submitted by 2379 different people, for a total cost of £ 4825.20 (including service fees, VAT, and the pre-test round). We used these answers to compile crowd-enVENT.

For the validation phase, we launched 6 rounds. Answers were collected with the software SoSciSurvey[1] which provides the possibility to create a questionnaire dynamically, with different texts for each participant. We estimated the completion time of a questionnaire to 8 minutes, and set the reward to £ 1 per participant. We further encouraged participants to follow the instructions with a bonus of £ 5 for the 5% best performing respondents (i.e., 60 crowdworkers whose appraisal reconstruction was the closest to the generation ratings). As we approved 1217 submissions, constructing the validation side of crowd-enVENT costed £ 2188.09 (VAT, service fees and bonus included).

Table A.3 reports an overview of the participants and the cost involved in both stages. In the generation block of the table (Phase 1), we indicate the strategy used in the text production task at each round:

- Strategy 0: participants were free to write any event of their choice.
- Strategy 1: they were asked to recount an event special to their lives.
- Strategy 2: they were shown the list of topics to avoid, described in Table 3.1 (Page 91).

The column "Workers" reports the number of different participants accepted in each round, hence in the row "$\sum$" is the total number of (unique) annotators whose answers entered the corpus, with the exception of the writers who contributed to round 1 (i.e., a pretest that we do not include in crowd-enVENT). Since the same writer could participate in multiple rounds, the sum of workers across the generation rounds exceeds 2379. In the validation block of the table

[1] https://www.soscisurvey.de.

		Phase 1		Phase 2	
Rounds	Strategies	Cost (£)	Workers	Cost (£)	Workers
1	0	156.1	–*	84	20
2	0	154.7	111	1474.1	1048
3	0	870.1	526	167.99	120
4	1 2	571.2 552.3	476	36.4	25
5	1 2	917.8 858.2	846	4.2	3
6	2	616.7	349	1.4	1
7	2	102.9	81		
8	2	10.5	13		
9	2	14.7	15		
$\sum$		4825.2	2379	1768.09	1217

Table A.3: Overview of study details in crowd-enVent, for each phase and round separately (with the relative text variability induction Strategies for generation), and after data aggregation. The answer of workers marked with an asterisk are excluded from the final corpus.

(Phase 2), $\sum$ = £ 1768.09 refers to the cost prior to releasing the bonus, which amounted to £ 420 (i.e., £ 300 for the extra reward to 60 people + commission charges).

1.4 crowd-enVent: Details on the Data Collection Questionnaires

The questionnaires in the generation and the validation phases of crowd-enVent were formulated in a comparable manner. Table A.4 makes their differences transparent, showing the templates across the two phases, and across the multiple rounds in the generation phase.

Note that some workers skipped the demographics- and personality-related portion of the survey, which had to be completed for them to be rewarded. We later allowed them to answer those questions in a separate form, containing only such questions.

	Question/Text	Value
Gx	**Study on Emotional Events**. Dear participant, Thanks for your interest in this study. We aim at understanding your evaluation of events in which you either felt a particular emotion or did not feel any. Further, we will ask you some demographic and personality-related information.The study should take you 4 minutes, and you will be rewarded with £ 0.50. Your participation is voluntary. You have to be at least 18 years old and a native speaker of English. Feel free to quit at any time without giving a reason (note that you won't be paid in this case). You can take this survey multiple times. You are also welcome to participate to the other versions of the survey that we published on Prolific, in which we ask you for your experience with different emotions. Note that towards the end of this survey, you will find a small set of questions that you only need to answer the first time you participate (which will save you time if you'll work on the other survey variants). The data we collect via Google forms will be used for research purposes. It will be made publicly available in an anonymised form. We will further write a scientific paper publication about this study which can include examples from the collected data (also in anonymous form). Nevertheless, please avoid providing information that could identify you (such as names, contact details, etc.). This study is funded by the German Research Foundation (DFG, Project Number KL 2869/1-2). Principle Investigator of this study: Dr. Roman Klinger, University of Stuttgart (Germany). Responsible and contact person: Enrica Troiano, University of Stuttgart (Germany). For any information, contact us at enrica.troiano@ims.uni-stuttgart.de	—

	Question/Text	Value
V	**Study on Emotional Events**. Dear participant, Thanks for your interest in this study. In a previous survey, people described events that might have triggered a particular emotion in them, and they answered some questions about those events. We now ask you to evaluate such events. You will read 5 brief event descriptions. For each of them, you will be asked the same questions that were answered by the event experiencers in the previous survey. Your task is to answer the same way as they did. Participants who are able to answer most similarly to the original authors will get a bonus of £5. We reward this bonus to the best 5% of participants. We will also ask you some demographic and personality-related information. There, your task is to provide information about yourself, and not about the author of the texts. The study should take you 8 minutes, and you will be rewarded with £1. Your participation is voluntary. You have to be at least 18 years old and a native speaker of English. Feel free to quit at any time without giving a reason (note that you won't be paid in this case). The data we collect will be used for research purposes. It will be made publicly available in an anonymised form. We will further write a scientific paper publication about this study which can include examples from the collected data (also in anonymous form). This study is funded by the German Research Foundation (DFG, Project Number KL 2869/1-2). Principle Investigator of this study: Dr. Roman Klinger, University of Stuttgart (Germany). Responsible and contact person: Enrica Troiano, University of Stuttgart (Germany). For any information, contact us at enrica.troiano@ims.uni-stuttgart.de	—
	I confirm that I have read the above information, meet the prerequisites for participation and want to participate in the study.	Yes/No
	Preliminary Questions. Please insert your ID as a worker on Prolific. Do you feel any of the following emotions right now, just before starting this survey? 1 means "not at all", 5 means "very intensely" [anger; boredom; disgust; fear; guilt; joy; pride; relief; sadness; shame; surprise; trust]	 Text Matrix with items [1–5]

	Question/Text	Value
GEx	**This study is about the emotional experience of E**. You will be asked to describe a concrete situation or an event which provoked this feeling in you and for which you vividly remember both the circumstance and your reaction. After that, you will be asked further information regarding such emotional experience, by indicating how much you agree with some statements on a scale from 1 to 5. Note: If you participated in our studies before, please describe a different situation now. We cannot accept an answer related to the same event you already told us about, even if you used different words. Further, we will not accept answers if they are not descriptions of events, like "I can't remember" or "I do not have that feeling".	—
GNx	**This study is about an experience you had, which did not involve you emotionally.** You will be asked to describe a concrete situation or an event which did not provoke any particular feeling in you and for which you vividly remember both the circumstance and your reaction. After that, you will be asked further information regarding such experience, by indicating how much you agree with some statements on a scale from 1 to 5. Note: If you participated in our studies before, please describe a different situation now. We cannot accept an answer related to the same event you already told us about, even if you used different words. Further, we will not accept answers if they are not descriptions of events, like "I can't remember" or "I always have feelings".	—
V	**Put yourself in the shoes of other people.** You will read five texts. These texts describe events that occurred in the life of their authors. Don't be surprised if they are not perfectly grammatical, or if you find that some words are missing. For each event, you will assess if it provoked an emotion in the experiencer, and if so, what emotion that was. Moreover, you will be asked how you think the experiencer assessed the event: you will read some statements and indicate how much you agree with each of them on a scale from 1 to 5. The writers of these texts have answered these questions in a previous survey. Your goal now is to guess the answer given by the writers as closely as possible.	—
GEx	**Recall an event that made you feel E.** Recall an event that made you feel E in the past.	—

	Question/Text	Value
GNx	**Recall an event that did not make you feel any emotion in the past.**	—
	It could be an event of your choice, or one which you might have experienced in one of the following areas: health, career, finances, community, fun/leisure, sports, arts, personal relationships, travel, education, shopping, learning, food, nature, hobbies, work... Please describe the event by completing the sentence below, including event details or write multiple sentences if this helps to understand the situation.	—
G1	The event should be special to you, or one which you think the other participants of this survey are unlikely to have experienced. It does not need to be an extraordinary event: it should just tell something about yourself.	—
G2	NOTE: We already collected many answers related to [OFF-LIMITS]. Please recount an event which does not relate to any of these: we need events which are as diverse as possible!	—
GEx	Please complete the sentence: I felt E when/because/...	Text
GNx	Please complete the sentence: I felt NO PARTICULAR EMOTION when/because/...	Text
V	What do you think the writer of the text felt when experiencing this event? [anger; boredom; disgust; fear; guilt; joy; pride; relief; sadness; shame; surprise; trust; no emotion]	single choice
V	How confident are you about your answer?	1…5
Gx	How long did the event last? [seconds; minutes; hours; days; weeks]	single choice
V	How long do you think the event lasted? [seconds; minutes; hours; days; weeks]	single choice
GEx	How long did the emotion last? [seconds; minutes; hours; days; weeks]	single choice
GNx	How long did the emotion last (if you had any)? [seconds; minutes; hours; days; weeks; I had none]	single choice
V	How long do you think the emotion lasted (if the experiencer had any)? [seconds; minutes; hours; days; weeks; this event did not cause any emotion]	single choice
Gx	How intense was your experience of the event?	1…5
V	How intense do you think the emotion was?	1…5
Gx	How confident are you that you recall the event well?	1…5

	Question/Text	Value
Gx	**Evaluation of that experience**. Think back to when the event happened and recall its details. Take some time to remember it properly. How much do these statements apply? (1 means "Not at all" and 5 means "Extremely")	—
V	**Evaluation of that Experience**. Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that event was perceived. How much do these statements apply? (1 means "I don't agree at all" and 5 means "I completely agree")	
	The event was sudden or abrupt.	1. . . 5
Gx	The event was familiar.	1. . . 5
V	The event was familiar to its experiencer.	1. . . 5
Gx	I could have predicted the occurrence of the event.	1. . . 5
V	The experiencer could have predicted the occurrence of the event.	1. . . 5
Gx	The event was pleasant for me.	1. . . 5
V	The event was pleasant for the experiencer.	1. . . 5
Gx	The event was unpleasant for me.	1. . . 5
V	The event was unpleasant for the experiencer.	1. . . 5
Gx	I expected the event to have important consequences for me.	1. . . 5
V	The experiencer expected the event to have important consequences for him/herself.	1. . . 5
	The event was caused by chance, special circumstances, or natural forces.	1. . . 5
Gx	The event was caused by my own behavior.	1. . . 5
V	The event was caused by the experiencer's own behavior.	1. . . 5
	The event was caused by somebody else's behavior.	1. . . 5
Gx	I anticipated the consequences of the event.	1. . . 5
V	The experiencer anticipated the consequences of the event.	1. . . 5
Gx	I expected positive consequences for me.	1. . . 5
V	The experiencer expected positive consequences for her/himself.	1. . . 5
	The event required an immediate response.	1. . . 5
Gx	I was able to influence what was going on during the event.	1. . . 5
V	The experiencer was able to influence what was going on during the event.	1. . . 5
Gx	Someone other than me was influencing what was going on.	1. . . 5
V	Someone other than the experiencer was influencing what was going on.	1. . . 5

	Question/Text	Value
	The situation was the result of outside influences of which nobody had control.	1...5
Gx	I anticipated that I would easily live with the unavoidable consequences of the event.	1...5
V	The experiencer anticipated that he/she could live with the unavoidable consequences of the event.	1...5
Gx	The event clashed with my standards and ideals.	1...5
V	The event clashed with her/his standards and ideals.	1...5
	The actions that produced the event violated laws or socially accepted norms.	1...5
Gx	I had to pay attention to the situation.	1...5
V	The experiencer had to pay attention to the situation.	1...5
Gx	I tried to shut the situation out of my mind.	1...5
V	The experiencer wanted to shut the situation out of her/his mind.	1...5
Gx	The situation required me a great deal of energy to deal with it.	1...5
V	The situation required her/him a great deal of energy to deal with it.	1...5
V	**Have you ever experienced an event similar to the one described?**	
	I experienced a similar event before.	1...5
Gx	**Is this the first time you participate in one of our emotional-event recollection studies?** We would like to know a bit more about you now. We have multiple similar studies on Prolific, all called ″Recollection of an emotion-inducing experience″, with the word ″emotion″ being replaced by an actual emotion name. When you participate in more than one of these studies, you only need to answer the following questions once. If this is the first time you participate, please answer them (otherwise we won't be able to approve your contribution), later you will skip this step. [Yes, first time, I will answer the following questions.; No, I participated before and answered the next set of questions.]	single choice
V	**Is this the first time you participate in our event evaluation studies?** If yes, you need to answer the following questions (otherwise we won't be able to approve your contribution). If no, you can skip them. [Yes, first time, I will answer the following questions.; No, I participated before and answered the next set of questions.]	single choice

	Question/Text	Value
Gx	**Demographic and Personality-related Questions**. As a last step, we ask you to answer some questions about yourself. Note: if you take one of our studies in the future, you won't fill in these sections again; if this is your first time and don't provide such information, we won't be able to reward you.	—
	How old are you?	int
	With which gender do you identify? Female; Male; Gender Variant/Non-Conforming; Prefer not to answer]	single choice
	What is the highest level of education you completed? [No formal qualifications; Secondary education; High school; Undegraduate degree (BA/BSc/other); Graduate degree (MA/MSc/MPhil/other); Doctorate degree (PhD/other); Don't know/ not applicable]	single choice
	With which of the following ethnic groups do you identify the most? [Australian/New Zealander; North Asian; South Asian; East Asian; Middle Eastern; European; African; North American; South American; Hispanic/Latino; Indigenous; Prefer not to answer; Other...]	single choice
	Here are a number of personality traits that may or may not apply to you. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other. [Extraverted, enthusiastic; Critical, quarrelsome; Dependable, self-disciplined; Anxious, easily upset; Open to new experiences, complex; Reserved, quiet; Sympathetic, warm; Disorganized, careless; Calm, emotionally stable; Conventional, uncreative]	Matrix with items [1. . . 7]
Gx	**One Last Question**. Please be assured that your answer will in no way influence how we treat your submission (you will be rewarded, if you properly followed our instructions). Did you actually experience that event or did you make it up to? [The event really happened in my life.; I never experienced that event, but I really imagined how it would make me feel.]	single choice

Table A.4: Template of the questionnaires for crowd-enVENT. The first column specifies the phase in which the question has been asked. Gn: Generation (GE: prompted by an emotion E, GN: prompted by the label "no emotion") with text production strategy n (see above, Section 1.3), V: Validation. No specification means that it has been asked in all variants. For the list of [OFF-LIMITS] topics in n = 2, refer to Table 3.1, Page 91.

1.5 deISEAR and enISEAR: Corpus Generation and Labeling

Restricting the countries of origin of the participants was crucial for data quality. To built deISEAR, we targeted Figure-Eight contributors from Germany and Austria. The English experiment was restricted to United Kingdom and Ireland. This prevented a substantial number of non-native participants who are proficient users of machine translation services from submitting answers. As a quality check, we required all workers to be level-3 contributors, i.e., the most experienced ones, who reached the highest accuracy in previous Figure-Eight jobs. These laypeople received no training from us, while participants of ISEAR were directly instructed by the experimenters. We aimed at adapting their questionnaire to a crowdsourcing framework, by formulating the task of sentence generation as one of sentence completion (e.g. *"Ich fühlte Freude, als/weil/..."*, *"I felt joy when/because ..."*). Preliminary experiments showed that people provided more coherent and grammatically correct sentences than when they were presented with a faithful translation of the original survey. The generation task was published in two slices (November/December 2018 and January 2019).

Phase 1 involved 121 English jobs and 116 German jobs after (manually) filtering out unacceptable answers (e.g., nonsensical items), totalling 2002 tasks (hits). The two languages required a diverse amount of jobs because of these ungrammatical and nonsensical descriptions that we discarded. In the second phase, 34 jobs were launched for English and 23 for German. This way we collected 5005 annotations for each language (i.e., 5 annotations per description). Overall, data generation and validation were finalized in three months. For each generated description, we paid 0.15 $; also for a validation task we paid 0.15 $, figuring that only annotating texts rather than producing them would take a comparable amount of time. The total cost was 300 $ for Phase 1, and 150 $ for Phase 2.

The top portion of Table A.5 illustrates the template instructions presented to the annotators for sentence generation (Phase 1), and the bottom shows a preview of the task itself. The corresponding survey of Phase 2 is presented in Table A.6.

Overview This study is about the emotional experience of E. Please recall a situation which provoked this feeling in you and for which you vividly remember both the circumstance and your reaction. Your responses will remain completely anonymous.

Rules Please complete the sentence below by describing a circumstance in you life that made you feel E.

I felt E because/when/that

- Responses that are not descriptions of events or situations are rejected and will not be paid. For example, "I can't remember" or "I do not have that feeling" are not acceptable answers.
- You will then answer three multiple-choice questions about this emotional experience.

Complete the sentence by describing a situation or event – in as much detail as possible – in which you felt E.

I felt E because/when/that

When did this happen?

○ days ago
○ weeks ago
○ months ago
○ years ago

How long did you feel the emotion?

○ a few minutes
○ an hour
○ several hours
○ a day or more

How intense was this feeling?

○ not very
○ moderately intense
○ intense
○ very intense

You are

○ female
○ male
○ other

Table A.5: Generation survey template for enISEAR (Phase 1). E is an emotion label placeholder. Top: Overview and instructions that introduced the task. Bottom: Actual task to complete.

Overview In an experiment we asked participants to describe emotional situations. Your task now is to guess which emotion was felt.

> *I felt . . . because I received more holidays than I thought I would get, so I could spend more time on my hobbies.*

Which emotion, do you think, did the experiencer feel?
○ anger
○ disgust
○ fear
○ guilt
○ joy
○ sadness
○ shame

Table A.6: Preview of the emotion validation task for enISEAR (Phase 2): overview (top) and task to complete (bottom). The sentence in the grey box is an example coming from Phase 1.

1.6 deISEAR and enISEAR: Guidelines for the Event Type Analysis

For each enISEAR description sampled for qualitative analysis, we annotated the following boolean variables:

- About the event time:
  - Does the text describe a *general event*?
  - Does the text describe a *future event*?
  - Does the text describe a *past event*?
- About the realization of the emotion:
  - Is it an actual or a *prospective* emotion?
- About the embedding in a social environment:
  - Are other people or animals part of the event description; is it a *social* event description?

- About the consequences of the event:
    - Are there *self-consequences*?
    - Are there *consequences for others*?
- About the control of the writer:
    - Is the author presumably under *situational control*?
    - Does the author presumably have *self control/responsibility*?

2 Explaining Disagreements with Extralinguistic Factors

This sections reports the task description and guidelines of the annotation conducted in Chapter 4. Both were shown to the coders. Q1 corresponds to EMO in the chapter, Q2 corresponds to CONF, and Q3 to INT.

In Q1, annotators were required to give their immediate, personal impression with respect to the presence of an emotion. Before annotating the 700 sentences selected from COCA, we observed inter-annotator agreement on a pre-annotation trial. With 70 sentences, Cohen's κ for pairs of annotators was found to be satisfactory (.52, .60 and .43) and motivated us to complete the job on the rest of the sentences. The job was completed upon a compensation of 60 €.

Task Description (Approximate Duration: 3 hours)

In this annotation trial, you will assess if texts are emotional or neutral.

Neutral sentences are those which

1. bear no affective connotation.

Emotional sentences are those which either

2. describe an event, a concept or state of affairs to which you would associate an emotion;
3. have an emotion as a central component of their meaning.

Examples of 1. are:

I am wearing my mask.

She answered her phone.

A new deal was established between the parties.

The elections are over.

Examples of 2. are:

I saw my bestfriend.

She was being pretty arrogant to me.

A war started in Westeros.

The king found an old sausage under his bed.

Examples of 3. are:

I am so happy to see you.

She was bursting with arrogance.

And there she was, desperate for her family.

I couldn't stand the catering food, bleark!

Guidelines

You will be shown individual sentences and, for each of them, you will answer 3 questions (**Q1**, **Q2**, **Q3**). Go for your immediate reaction to the text – avoid over-assessments.

Q1 Given a sentence, please ask yourself: **is it emotional (E) or neutral (N)?** Type:

- **N**, if the text does not convey any emotion, like in Examples of 1.;
- **E**, if an emotion could be inferred from the text, like in Examples of 2., or an emotion is a central part of the text, like in Examples of 3.

Q2 Ask yourself: **how confident am I about my answer to Q1?** Give yourself a rating on a scale from 1 to 3. Indicate if you are

- **1**, not confident at all;
- **2**, confident;
- **3**, sure.

Q3 This question only applies if you answered 2 or 3 in **Q1**: in case the sentence expresses an emotion, **how strong is such emotion?** Assess the degree of its intensity on a scale from 1 to 3, where

- **1** is mild;
- **2** is intense;
- **3** is very intense.

Appendix B

Additional Analyses

1 Emotion Detection from Implicit Expressions

This section provides additional analyses for Chapter 3.

1.1 Descriptive Analysis of crowd-enVENT

Table B.1 shows the most frequent noun lemmata[1] as a proxy of the described topics. Besides reoccurring terms (e.g., family- and work-related ones), which are used to contextualize the events themselves, some words are more specific to certain emotions, and they indicate concepts that have a prototypical emotion meaning in the collective imagination, like "*spider*" and "*night*" for *fear*, "*birthday*" for *surprise*, "*degree*" and "*award*" for *pride*.

A frame-oriented overview of the semantics of crowd-enVENT is below in this Appendix, Section 4.3.

[1] Calculated via SpaCy v.3.2, `https://spacy.io/api/lemmatizer`.

Emotion	Most Frequent Nouns
Anger	work friend time partner car people child year day job husband family boyfriend son member school mother colleague week house daughter thing person ex
Boredom	work time hour day home job friend class room night meeting game week one thing house training task phone flight tv school lecture weekend traffic lot
Disgust	friend people man food dog work time child family day house person partner colleague car floor boyfriend street room parent job school night member cat
Fear	car night time friend day house dog year child work hospital road man people accident family dad spider son partner front job hour door way phone park life
Guilt	friend time child work money partner girlfriend day thing family brother school mother son sister relationship daughter year dog ex dad parent lot kid father
Joy	time friend year day child family boyfriend son job dog partner birthday birth baby work school life daughter car week room month wife song sister holiday
No Emo.	morning job time work day friend boyfriend year school car thing grocery today event life situation shop tv task shopping people partner family college
Pride	work job year son time school daughter university friend day degree award team lot week child game student class college exam family company result
Relief	time day work job test house year week friend daughter result car surgery school month dog exam cancer university partner money home health son night
Sadness	friend year time family job dog dad day week child month boyfriend sister mum life parent daughter cat work husband school house home thing people
Shame	friend work school money day time parent front family test thing people sister member exam situation sex lot dad class child year wife store partner job
Surprise	friend birthday year time job party boyfriend work sister partner gift car wife week parent girlfriend month money day trip person husband house college
Trust	friend partner time boyfriend husband work life secret family car relationship people job doctor day girlfriend situation hospital colleague money year person

Table B.1: Most frequent nouns for each prompting emotion category in crowd-enVENT, sorted by frequency.

		Temp. Dist.				Intensity				Duration				Gender		
	Emotion	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	.19	.12	.13	.13	.05	.10	.18	.15	.07	.16	.18	.18	.14	.14	0
	Disgust	.16	.18	.17	.08	.21	.20	.13	.10	.31	.20	.04	.01	.14	.15	0
	Fear	.10	.16	.15	.16	.07	.09	.16	.18	.16	.18	.14	.10	.14	.16	0
	Guilt	.15	.13	.12	.16	.14	.22	.15	.08	.13	.16	.20	.10	.15	.12	0
	Joy	.17	.15	.12	.14	.04	.07	.16	.20	.05	.10	.20	.23	.14	.16	1
	Sadness	.12	.13	.17	.15	.05	.12	.12	.21	.05	.05	.13	.31	.14	.14	0
	Shame	.10	.14	.15	.17	.43	.21	.11	.07	.23	.15	.11	.06	.15	.12	0
English	Anger	.18	.16	.13	.11	.12	.12	.15	.16	.13	.13	.16	.15	.15	.14	0
	Disgust	.23	.14	.11	.10	.16	.18	.12	.13	.28	.15	.11	.07	.13	.15	0
	Fear	.08	.16	.19	.15	.03	.10	.18	.17	.22	.16	.15	.07	.16	.13	0
	Guilt	.13	.14	.14	.15	.33	.18	.14	.07	.11	.22	.12	.14	.14	.15	0
	Joy	.13	.14	.16	.14	.03	.09	.15	.20	.06	.07	.19	.20	.14	.14	0
	Sadness	.16	.14	.16	.12	.13	.16	.12	.16	.07	.12	.10	.23	.15	.14	0
	Shame	.09	.12	.10	.21	.21	.18	.13	.11	.12	.14	.17	.14	.13	.15	0

Table B.2: Statistics on deISEAR and enISEAR normalized by column. The unnormalized counts are shown in Chapter 3, Table 3.3. Temp. Dist.: temporal distance.

1.2 Descriptive Analysis of de(/en)ISEAR

Table B.2 and Table B.3 present a compact description of our multilingual resources, normalizing the counts by column and by row blocks, respectively.

Table B.2 highlights differences in the distribution of emotions across different values of *temporal distance*, *intensity*, *emotion duration*, and annotators' *gender*. We see for instance that *shame* is outstanding in English for long-distant events, while *anger* and *disgust* (depending on the language) are more dominant in events that happened a few days prior to description production. For *intensity*, the distribution of scores across emotions is the most unbalanced with the label "not very"; for *emotion duration*, *disgust* is the prevailing emotion among those which lasted only a few minutes, while it is the less frequent among those which persisted one or multiple days. The exact opposite holds for *joy* and *sadness*, which appear to be more durable states.

Table B.3 highlights differences in the distribution of extralinguistic labels across different emotions. A few commonalities emerge between the two languages. The majority of descriptions are referred to remote emotion episodes. Moreover, *anger-*, *fear-*, *joy-* and *sadness*-related de-

		Temp. Dist.				Intensity				Duration				Gender		
	Emotion	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	.32	.17	.22	.29	.02	.17	.47	.34	.16	.20	.27	.36	.78	.22	0
	Disgust	.27	.27	.29	.17	.08	.36	.34	.22	.66	.26	.06	.02	.77	.23	0
	Fear	.17	.22	.26	.34	.03	.17	.41	.40	.35	.22	.22	.21	.76	.24	0
	Guilt	.25	.19	.21	.35	.06	.40	.38	.17	.29	.20	.30	.21	.81	.19	0
	Joy	.28	.21	.20	.31	.01	.13	.42	.44	.10	.13	.29	.48	.75	.24	.01
	Sadness	.20	.18	.29	.32	.02	.22	.30	.46	.11	.06	.19	.64	.79	.21	0
	Shame	.17	.20	.25	.38	.17	.39	.29	.15	.50	.20	.17	.13	.81	.19	0
English	Anger	.31	.20	.17	.31	.06	.24	.34	.36	.21	.16	.25	.38	.43	.57	0
	Disgust	.40	.17	.15	.28	.08	.36	.26	.30	.46	.19	.17	.18	.40	.60	0
	Fear	.13	.20	.25	.41	.01	.21	.40	.38	.36	.20	.24	.19	.46	.54	0
	Guilt	.23	.17	.19	.41	.17	.36	.30	.16	.18	.27	.20	.35	.41	.59	0
	Joy	.22	.17	.22	.39	.01	.19	.34	.46	.10	.09	.30	.51	.42	.58	0
	Sadness	.28	.17	.22	.34	.07	.31	.27	.35	.12	.15	.16	.57	.43	.57	0
	Shame	.15	.15	.13	.57	.11	.36	.29	.24	.20	.17	.27	.35	.40	.60	0

Table B.3: Statistics on deISEAR and enISEAR normalized by partial row. The unnormalized counts are shown in Chapter 3, Table 3.3. Temp. Dist.: temporal distance.

scriptions are mostly about events which caused very intense affective states. For *emotion duration*, most occurrences of *anger* and *sadness* lasted longer than one day both in German and English, while *fear* episodes are more short-termed, similar to *disgust*.

1.3 Event-type Analysis

The event-type analysis of enISEAR and deISEAR targeted 385 items per language (55 descriptions per emotion). Table 3.4 in Chapter 3 showed the counts of instances associated to the psychology-motivated labels across the seven emotions. While that described the distribution of labels by emotion, here we expand the discussion to the extralinguistic information collected in Phase 1. Table B.4 distributes the raw counts across the annotation values. It should be noticed that the random descriptions used for this analysis were not balanced with respect to their values of each variable. For this reason, Table B.5 reports relative counts (i.e. counts of descriptions normalized by the number of instances within the labels "days" (D), "weeks" (W), "months" (M) etc.).

Some regularities can be observed cross all columns of Table B.5.

		Temp. Dist.				Intensity				Duration				Gender		
	Dimension	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	General Event	2	3	1	2	0	1	4	3	4	0	1	3	6	2	0
	Future Event	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0
	Past Event	98	76	101	101	22	92	141	121	121	66	83	106	287	89	0
	Prospective	3	2	2	0	0	3	3	1	2	2	1	2	5	2	0
	Social	55	41	53	51	13	43	80	64	70	32	42	56	152	48	0
	Self conseq.	54	45	70	67	15	52	94	75	74	36	59	67	176	60	0
	Conseq. oth.	42	30	34	41	10	35	54	48	52	25	28	42	110	37	0
	Situat. control	17	13	18	18	2	17	29	18	21	10	14	21	56	10	0
	Own resp.	53	37	63	55	11	57	76	64	68	40	45	55	160	48	0
	Sum	226	171	242	234	51	208	341	273	291	145	191	246	666	207	0
English	General Event	6	2	2	1	2	2	5	2	5	2	1	3	3	8	0
	Future Event	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Past Event	88	61	73	152	23	104	122	125	76	74	85	139	155	219	0
	Prospective	3	4	3	5	0	2	8	5	5	4	3	3	7	8	0
	Social	73	51	56	107	14	72	94	107	49	52	66	120	103	184	0
	Self conseq.	46	30	34	90	14	57	71	58	52	32	47	69	89	111	0
	Conseq. oth.	51	38	39	73	8	49	69	75	30	47	40	84	73	128	0
	Situat. control	15	17	16	42	12	30	25	23	21	19	17	33	40	50	0
	Own resp.	50	36	47	89	20	71	80	51	57	50	53	62	104	118	0
	Sum	244	178	197	407	70	283	352	321	219	206	227	374	419	607	0

Table B.4: Event-type analysis. Raw counts of the labels manually assigned to a subset of enISEAR and deISEAR, across the extralinguistic information collected in Phase 1. Temp. Dist.: temporal distance. Self conseq./Conseq. oth.: consequences for the self/others. Own resp.: own responsibility/own control.

Events that involved a purposeful participation of their experiencer are a minority in both languages (Sit. control). Approximately 50% of the descriptions mention individuals other than the writer (Social). This proportion, however, is higher for English than for German.

Events linked to consequences for the self mostly come from the German sample (Self conseq.). In German, such types of events are recalled more frequently than events that had consequences on others (Conseq. oth.). The opposite is true for English: English authors often wrote about events that affected others. This holds irrespective of the *temporal distance*, the *intensity*, the *emotion duration* and the *gender* of the experiencer. Exceptions are English descriptions of emotion facts which only lasted a few minutes, and which appear to bring consequences for the self more than for others (Self conseq. and Conseq. oth. in column "min"). The personal responsibility or control of events is

		Temp. Dist.				Intensity				Duration				Gender		
	Dimension	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	General Event	.02	.04	.01	.02	0	.01	.03	.02	.03	0	.01	.03	.02	.02	0
	Future Event	0	0	.01	0	0	0	.01	0	0	0	.01	0	0	0	0
	Past Event	.98	.96	.98	.98	1	.99	.97	.98	.97	1	.98	.97	.98	.98	0
	Prospective	.03	.03	.02	0	0	.03	.02	.01	.02	.03	.01	.02	.02	.02	0
	Social	.55	.52	.51	.50	.59	.46	.55	.52	.56	.48	.49	.51	.52	.53	0
	Self conseq.	.54	.57	.68	.65	.68	.56	.64	.60	.59	.55	.69	.61	.60	.66	0
	Conseq. oth.	.42	.38	.33	.40	.45	.38	.37	.39	.42	.38	.33	.39	.37	.41	0
	Situat. control	.17	.16	.17	.17	.09	.18	.20	.15	.17	.15	.16	.19	.19	.11	0
	Own resp.	.53	.47	.61	.53	.50	.61	.52	.52	.54	.61	.53	.50	.54	.53	0
English	General Event	.06	.03	.03	.01	.08	.02	.04	.02	.06	.03	.01	.02	.02	.04	0
	Future Event	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Past Event	.94	.97	.97	.99	.92	.98	.96	.98	.94	.97	.99	.98	.98	.96	0
	Prospective	.03	.06	.04	.03	0	.02	.06	.04	.06	.05	.03	.02	.04	.04	0
	Social	.78	.81	.75	.70	.56	.68	.74	.84	.60	.68	.77	.85	.65	.81	0
	Self conseq.	.49	.48	.45	.59	.56	.54	.56	.46	.64	.42	.55	.49	.56	.49	0
	Conseq. oth.	.54	.60	.52	.48	.32	.46	.54	.59	.37	.62	.47	.59	.46	.56	0
	Situat. control	.16	.27	.21	.27	.48	.28	.20	.18	.26	.25	.20	.23	.25	.22	0
	Own resp.	.53	.57	.63	.58	.80	.67	.63	.40	.70	.66	.62	.44	.66	.52	0

Table B.5: Event-type analysis. Counts are normalized by instances with a particular annotation value, e.g., the count in the cell "Time General"–"D" is normalized by the number of all instances with the associated value D (temporal distance of days). Temp. Dist.: temporal distance. Self conseq./Conseq. oth.: consequences for the self/others. Own resp.: own responsibility/own control.

consistent across all columns in the German sample. Instead, in English we observe some marked differences. Emotions with a low intensity (column "NV") followed an event which was directly triggered by their experiencer, but very intense emotions are less frequently associated to responsibility (column "VI"). Lastly, shorter events ("min") imply the responsibility dimension more than long ones (≥d).

1.4 Inter-Annotator Agreement on de(/en)ISEAR

Section 5.1.2 in Chapter 3 discussed the agreement reached by different subsets of annotators *for each generation label* of our multilingual resources. We report the relative counts in Table B.6.

We extend the analysis in Table B.7, summing over the prompting emotions. This table shows the inter-annotator agreement of Phase 2

	German					English				
Emotion	≥1	≥2	≥3	≥4	=5	≥1	≥2	≥3	≥4	=5
Anger	.94	.87	.75	.57	.36	.96	.90	.78	.62	.41
Disgust	.97	.94	.91	.87	.64	.83	.71	.59	.53	.37
Fear	.94	.87	.76	.69	.55	.95	.92	.87	.81	.60
Guilt	.96	.88	.71	.47	.22	.96	.91	.87	.62	.31
Joy	.99	.99	.99	.98	.95	1	1	1	1	.96
Sadness	.92	.86	.79	.68	.53	.98	.93	.92	.81	.68
Shame	.90	.76	.60	.46	.29	.81	.64	.45	.29	.16
Sum	6.62	6.17	5.51	4.71	3.53	6.48	6.01	5.47	4.69	3.49

Table B.6: Relative agreement counts on de(/en)ISEAR.

with respect to the meta-information given by the participants of Phase 1, i.e., *gender, intensity, emotion duration* and *temporal distance* under the column "Labels". Numbers represent the count of descriptions (within a corpus, and not within a generation label), for which the annotation label is the same as the generation label. The table can be read as follows: 177 descriptions from deISEAR, which were labeled as very intense ("VI") by Phase 1 annotators, were then labeled by five Phase 2 annotators with their original prompting emotion; 506 instances provided by female annotators for enISEAR were labeled by at least two Phase 2 annotators with their original prompting emotion, and so on.

The corresponding analysis in Chapter 3 is reported in Table 3.9. There, the maximum value that each cell can reach is 143, i.e., the total number of descriptions prompted by a specific emotion. Here, the maximum value varies by cell, because each meta-data label is assigned to a different number of descriptions.[2] Accordingly, higher counts do not necessarily indicate stronger agreement.

1.5 Automatic Classification on enISEAR and deISEAR

Table B.8 shows the results of the Maximum Entropy classifier across all ground truth emotions.

[2] For the distribution of meta-data labels over the descriptions, refer to Section 4.2 in Chapter 3.

		German					English				
	Labels	≥1	≥2	≥3	≥4	=5	≥1	≥2	≥3	≥4	=5
When	D	226	157	184	209	226	229	211	189	161	115
	W	197	184	169	143	108	168	152	137	112	79
	M	229	215	198	174	125	177	165	154	138	109
	Y	295	275	237	200	161	353	331	302	259	196
Length	min	291	275	245	213	145	223	208	185	162	115
	h	173	162	151	127	99	162	145	130	106	74
	>h	205	188	164	139	103	210	197	178	158	118
	≥d	278	258	228	195	158	332	309	289	244	192
Intense	NV	52	46	38	32	18	74	69	61	51	31
	M	241	224	194	162	113	264	240	217	185	128
	I	352	331	301	255	197	288	267	247	213	165
	VI	302	282	255	225	177	301	283	257	221	172
Gender	M	738	684	604	510	392	386	353	316	273	200
	F	208	198	183	163	112	541	506	466	397	299
	O	1	1	1	1	1	–	–	–	–	–

Table B.7: Full agreement information for de(/en)ISEAR.

	deISEAR						enISEAR					
Emotion	TP	FP	FN	P	R	F_1	TP	FP	FN	P	R	F_1
Anger	29	30	114	.49	.20	.29	27	32	116	.46	.19	.27
Disgust	65	57	78	.53	.45	.49	67	85	76	.44	.47	.45
Fear	70	77	73	.48	.49	.48	85	69	58	.55	.59	.57
Guilt	75	140	68	.35	.52	.42	79	161	64	.33	.55	.41
Joy	106	61	37	.63	.74	.68	94	43	49	.69	.66	.67
Sadness	63	31	80	. 67	.44	.53	70	29	73	.71	.49	.58
Shame	66	131	77	.34	.46	.39	49	111	94	.31	.34	.32
Micro	474	527	527	.47	.47	.47	471	530	530	.47	.47	.47

Table B.8: Details of classification results on deISEAR and enISEAR.

1.6 Experiencer-specific Appraisal Annotation on enISEAR

With an ancillary study conducted on enISEAR, we gathered post-annotations about the appraisals of all participants in a given event description. We also collected annotations about the emotions they likely felt in that situation. The study closed the gap between the discrete emotion annotation of enISEAR and the appraisal-based framework of crowd-enVENT.[3] In this section, I will refer to this upscaled version of enISEAR as x-enVENT, with "x" standing for experts, as we relied on annotators trained in-house, with detailed definitions for each appraisal dimension. Note that x-enVENT does not contain all 1001 descriptions of enISEAR, but only 683 items (due to time restrictions on the annotation activity). It further includes 37 texts that were extracted from other corpora; they are not considered in this analysis.

We utilized the 22 appraisals reported in Table B.9, many of which denote the same dimensions of crowd-enVENT. Exceptions are *exert*, *consider*, *understand*, *expectation discrepancy*, that were used on x-enVENT but not on crowd-enVENT (vice versa, *unpleasantness*, *event predictability*, *not consider* are in crowd-enVENT but not in x-enVENT). As the publication of this study (Troiano et al., 2022a) refers to them with a slightly different notation than the one presented in Chapter 3 (Page 80), the table details the names with which they were indicated and the definitions shown to the annotators. For consistency with the rest of the dissertation, I use the corresponding names of crowd-enVENT. The mapping is presented in the table as well.

[3] Note that crowd-enVENT does not contain emotions and appraisals associated to each event participant, but only to the writers.

Variable	Definition	Corresponding in crowd-enVENT
suddenness	the event was sudden or abrupt	*suddenness*
familiarity	the event was familiar for the experiencer	*familiarity*
pleasantness	the event was pleasant for the experiencer	*pleasantness*
goal relevance	the event was important or relevant for experiencer's goals	*goal relevance*
situational responsibility	the event was caused by chance or special circumstances	*situational responsibility*
self responsibility	the event was caused by experiencer's own behaviour	*own responsibility*
other responsibility	the event was caused by somebody else's behaviour	*other responsibility*
outcome probability	the experiencer could anticipate the consequences of the event	*anticip. conseq.*
goal conduciveness	the event itself was positive or it had positive consequences for the experiencer	*goal support*
urgency	the event required an immediate response from the experiencer	*urgency*
self control	the experiencer had the capacity to affect the event	*own control*
other control	someone or something other than the experiencer was influencing what was going on	*others' control*
situational control	the situation was the result of outside influences of which nobody had control	*situational control*
adjustment check	the experiencer anticipated that they could live with the consequences of the event	*accept. conseq.*
internal check	the event clashed with the experiencer's ideals and standards	*internal standards*
external check	the event violated laws or social norms.	*external norms*
attend	the experiencer had to pay attention to the situation	*attention*
effort	the situation required the experiencer a great deal of energy	*effort*

exert	the experiencer felt they needed to exert themselves to handle the event	–
consider	the experiencer wanted to consider the situation	–
understand	the experiencer understood what was happening	–
expectation discrepancy	the experiencer did not expect that the event would occur	–

Table B.9: Names and definitions of appraisal variables used in the experiencer-specific study of enISEAR, and their mapping to the appraisal dimensions of crowd-enVENT.

Having appraisal scores assigned to different semantic roles, we are interested in two aspects, treated separately in the paragraphs below. One is the relationship between the event evaluations of the two types of entities in x-enVENT, divided into the narrating voice (i.e., the writer) and others. The second aspect is the comparison of the distribution of appraisal assigned to the writers between x-enVENT and crowd-enVENT.

Between-Entity Appraisals. If an entity perceives, for instance, *own responsibility* for an event, what is the appraisal of the other experiencer? For each (writer–other entity) pair in a text, we retrieve the scores of all appraisal pairs, where one element is the score (averaged across the answers of 4 coders) assigned to the writer and the other is given to the mentioned entity. Hence, we calculate Spearman's correlation for such appraisal combinations.

The resulting correlations are in Figure B.1. Appraisals that hold for the writers are positively correlated with the same dimensions for other entities (see diagonal). This, however, does not suggest that events are always similarly appraised by all participants, as many positive correlations can be found among diverse appraisal combinations. Examples are *internal standards–external norms* ($\rho = .68$), *expectation discrepancy-external check* (.45), and *others' control* (or *responsibility*)–*own control* (*responsibility*). The latter pair indicates that, often, one participant triggers the event and the other is subjected to it – but if an event is driven by external factors, it is so for both (see their *situational control/responsibility*). Among the negatively correlated dimensions, we notice *internal standards–pleasantness*, *anticipated consequences–urgency*, *external norms–accepted consequences*.

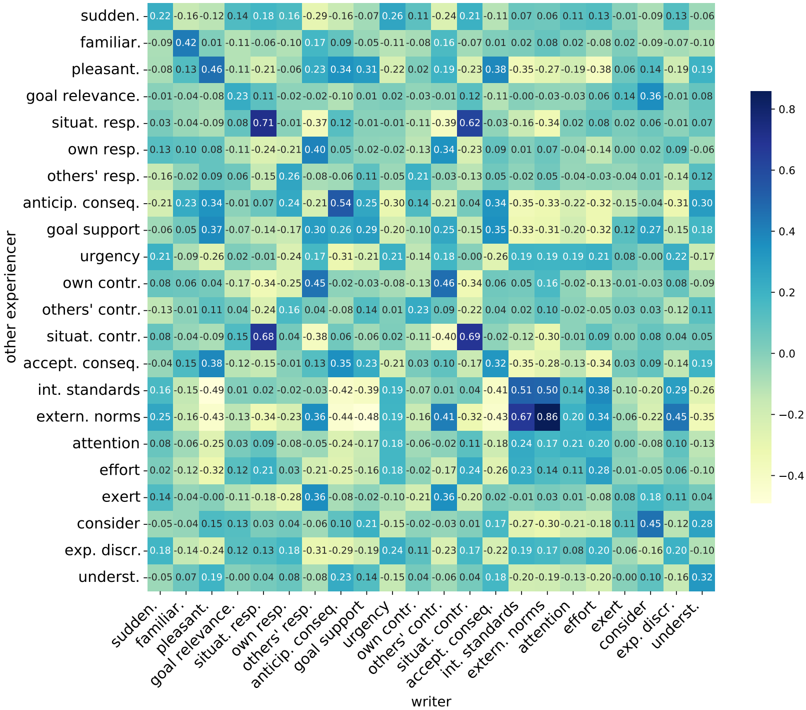
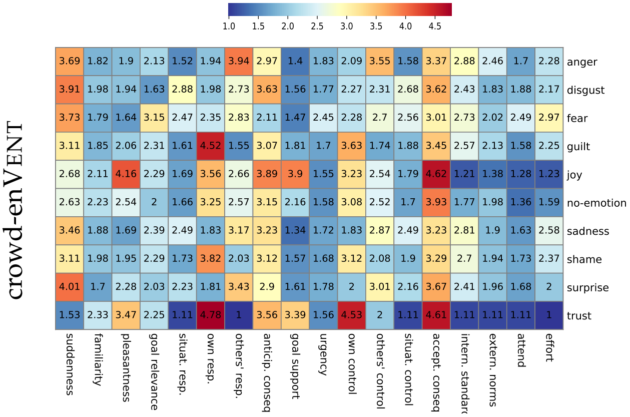


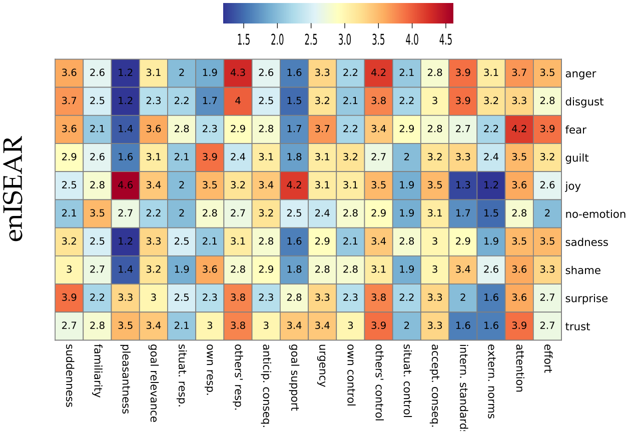
Figure B.1: Spearman's correlations between the writers' (columns) and other experiencers' (rows) appraisal scores.

Within-Experiencer Appraisal Distribution. Considering crowd-enVENT and x-enVENT, we observe the distribution of appraisal ratings attributed to the writers by the readers. Figure B.2 shows the average appraisal values across emotions (i.e., the inferred emotions, and not the prompting ones). It includes the intersection of each variable between the two corpora, i.e., we report only the emotions selected by both the validators of crowd-enVENT and of x-enVENT, and the appraisal dimensions present in the two studies.

The numbers in the cells of the two heatmaps are not immediately comparable because the judgments from which they are calculated concern different texts and were collected in different setups (crowd-enVENT via crowdsourcing, x-enVENT in-lab and with a handful of coders who annotated all texts). However, it is possible to find correspondences between patterns of appraisals.



suddenness	familiarity	pleasantness	goal relevance	situat. resp.	own resp.	others' resp.	anticip. conseq	goal support	urgency	own control	others' control	situat. control	accept. conseq	intern. standarc	extern. norms	attend	effort	
3.69	1.82	1.9	2.13	1.52	1.94	3.94	2.97	1.4	1.83	2.09	3.55	1.58	3.37	2.88	2.46	1.7	2.28	anger
3.91	1.98	1.94	1.63	2.88	1.98	2.73	3.63	1.56	1.77	2.27	2.31	2.68	3.62	2.43	1.83	1.88	2.17	disgust
3.73	1.79	1.64	3.15	2.47	2.35	2.83	2.11	1.47	2.45	2.28	2.7	2.56	3.01	2.73	2.02	2.49	2.97	fear
3.11	1.85	2.06	2.31	1.61	4.52	1.55	3.07	1.81	1.7	3.63	1.74	1.88	3.45	2.57	2.13	1.58	2.25	guilt
2.68	2.11	4.16	2.29	1.69	3.56	2.66	3.89	3.9	1.55	3.23	2.54	1.79	4.62	1.21	1.38	1.28	1.23	joy
2.63	2.23	2.54	2	1.66	3.25	2.57	3.15	2.16	1.58	3.08	2.52	1.7	3.93	1.77	1.98	1.36	1.59	no-emotion
3.46	1.88	1.69	2.39	2.49	1.83	3.17	3.23	1.34	1.72	1.83	2.87	2.49	3.23	2.81	1.9	1.63	2.58	sadness
3.11	1.98	1.95	2.29	1.73	3.82	2.03	3.12	1.57	1.68	3.12	2.08	1.9	3.29	2.7	1.94	1.73	2.37	shame
4.01	1.7	2.28	2.03	2.23	1.81	3.43	2.9	1.61	1.78	2	3.01	2.16	3.67	2.41	1.96	1.68	2	surprise
1.53	2.33	3.47	2.25	1.11	4.78	1	3.56	3.39	1.56	4.53	2	1.11	4.61	1.11	1.11	1.11	1	trust



suddenness	familiarity	pleasantness	goal relevance	situat. resp.	own resp.	others' resp.	anticip. conseq.	goal support	urgency	own control	others' control	situat. control	accept. conseq.	intern. standard	extern. norms	attention	effort	
3.6	2.6	1.2	3.1	2	1.9	4.3	2.6	1.6	3.3	2.2	4.2	2.1	2.8	3.9	3.1	3.7	3.5	anger
3.7	2.5	1.2	2.3	2.2	1.7	4	2.5	1.5	3.2	2.1	3.8	2.2	3	3.9	3.2	3.3	2.8	disgust
3.6	2.1	1.4	3.6	2.8	2.3	2.9	2.8	1.7	3.7	2.2	3.4	2.9	2.8	2.7	2.2	4.2	3.9	fear
2.9	2.6	1.6	3.1	2.1	3.9	2.4	3.1	1.8	3.1	3.2	2.7	2	3.2	3.3	2.4	3.5	3.2	guilt
2.5	2.8	4.6	3.4	2	3.5	3.2	3.4	4.2	3.1	3.1	3.5	1.9	3.5	1.3	1.2	3.6	2.6	joy
2.1	3.5	2.7	2.2	2	2.8	2.7	3.2	2.5	2.4	2.8	2.9	1.9	3.1	1.7	1.5	2.8	2	no-emotion
3.2	2.5	1.2	3.3	2.5	2.1	3.1	2.8	1.6	2.9	2.1	3.4	2.8	3	2.9	1.9	3.5	3.5	sadness
3	2.7	1.4	3.2	1.9	3.6	2.8	2.9	1.8	2.8	2.8	3.1	1.9	3	3.4	2.6	3.6	3.3	shame
3.9	2.2	3.3	3	2.5	2.3	3.8	2.3	2.8	3.3	2.3	3.8	2.2	3.3	2	1.6	3.6	2.7	surprise
2.7	2.8	3.5	3.4	2.1	3	3.8	3	3.4	3.4	3	3.9	2	3.3	1.6	1.6	3.9	2.7	trust

Figure B.2: Distribution of average appraisal values across the emotions assigned to crowd-enVENT and part of enISEAR (i.e., x-enVENT) by readers.

As an example, for both resources, *surprise* and *disgust* are high in *suddenness*, *trust* and *noemotion* in *familiarity*, *fear* in *goal relevance*, *surprise* and *sadness* are characterized by a strong *situational responsibility*. We conclude that, despite differences between the data under consideration, the underlying properties of emotions identified by the readers are similar.

2 Explaining Disagreements with Extralinguistic Factors

Section 3.2.2 in Chapter 4 discussed how being characterized by a property and sharing that property corresponds to changes in the agreement between validators. Taking a different perspective, here we observe if any of the properties under consideration has an impact on the correctness of emotion and appraisal reconstruction.

The procedure we use is similar to that with the V–V pairs, but it focuses on the agreement between generators and validators. First, we couple all study participants who respectively generated and validated the same texts, obtaining 6,600 generator–validator (G–V) pairs (each writer is coupled with 5 readers. We then filter them according to the various properties (using the same thresholds and criteria presented on Page 133), and only consider the intersection of the obtained subsets, like we did earlier. The difference here is that the properties in question are looked for only in the validator of each pair. Hence, we only form two groups for each factor, comprising pairs in which the validator has a property, and pairs in which the validator does not have it.

Table B.10 summarizes the results for the factors *gender*, *age*, *event familiarity* and *personality traits*. We see that *gender* does not significantly improve the validators' performance. Interestingly, validators unfamiliar with the described event achieve better appraisal reconstructions. Differences among agreements conditioned on *age* are small (1pp in acc.), but still significant for appraisal. Concerning *personality traits*, openness does not show any significant relation to agreement across measures and annotation variables, as opposed to conscientiousness, which shows a small but significant positive impact on all measures. For the other traits, the negative dimension (e.g., not extravert) improves appraisal reconstruction.

Table B.11 reports results obtained with the factors *education*, *ethnicity* and *current emotion state* (properties with no significant difference across variables are omitted). We did not find any substantial difference between the agreement achieved by groups of a specific *ethnicity*, but we found coders with a higher degree of *education* to reconstruct appraisals better. We further observe an effect of diverse *emotion states* concerning both appraisal and emotion judgments, to a various degree for different states.

In summary, better emotion and appraisal reconstructions are favored by specific properties.

			Agreement		
			Emotion		Appraisal
Factor	Validator Property	#Pairs	F_1	Acc.	RMSE
None	All data	6600	.49	.50	1.57
Gender	male	2286	.49	.50	1.56
	female	3019	.50	.51	1.57
Event Fam.	familiar	1386	.49	.51	*1.60
	unfamiliar	2099	.48	.49	*1.58
Age diff.	≤ 7	2076	.49	.50	*1.56
	> 7	3089	.49	.51	*1.58
Pers. Traits	open	2685	.49	.50	1.57
	not open	3000	.49	.50	1.57
	conscientious	3151	*.48	*.49	*1.57
	not conscientious	2589	*.50	*.51	*1.56
	extravert	2878	.49	.50	*1.58
	not extravert	2812	.50	.51	*1.56
	agreeable	2675	.49	.51	*1.58
	not agreeable	2930	.48	.49	*1.56
	emo. stable	2838	*.48	*.49	*1.57
	emo. unstable	2792	*.50	*.51	*1.56

Table B.10: *Gender, Event Familiarity, Age* and *Personality Traits* as conditions of agreement between generators and validators, shown as F_1 and Accuracy (for emotions) and average root mean square error (for appraisals). For each factor, the column "#Pairs" reports the size of a sample on which agreement is computed. Colored boxes indicate numbers that can be compared to each other. "*" indicates that they are significantly different, as found with 1000× bootstrap resampling, confidence level .95.

			Agreement		
			Emotion		Appraisal
Factor	Validator Property	#Pairs	F_1	Acc.	RMSE
Education	High school	820	.47	.49	$^{\dagger *}$1.60
	Ungraduate Degree	955	.49	.49	*1.57
	Graduate Degree	683	.49	.50	†1.55
Current State	anger	186	.48	.49	*1.64
	not anger	694	.49	.51	*1.57
	boredom	1444	*.51	.52	1.57
	not boredom	3166	*.48	.50	1.56
	disgust	126	*.46	.48	*1.69
	not disgust	444	*.50	.51	*1.58
	joy	2209	*.47	*.49	*1.59
	not joy	3081	*.52	*.53	*1.55
	pride	1675	*.47	*.48	*1.61
	not pride	3175	*.50	*.51	*1.55
	relief	1289	.51	.52	*1.62
	not relief	2731	.50	.51	*1.55
	sadness	660	.51	*.53	1.54
	not sadness	2050	.48	*.50	1.55
	surprise	440	*.44	.46	*1.65
	not surprise	1395	*.49	.50	*1.57
	trust	2574	.48	.50	*.57
	not trust	3046	.50	.50	*1.55

Table B.11: *Education* and *Current Emotion State* as conditions of agreement between generators and validators, shown as F_1 and Accuracy (for emotions) and average root mean square error (for appraisals). For each factor, the column "#Pairs" reports the size of a sample on which agreement is computed. Colored boxes indicate numbers that can be compared to each other. "*" indicates that they are significantly different, as found with 1000× bootstrap resampling, confidence level .95. *Current Emotion State* only reports emotions for which we found significant differences.

3 Is Emotion a "Style?"

Table B.12 reports challenging examples for the MT-based pipeline described in Chapter 5. Sentences corresponding to RQ1 show how the same sentence is transformed by the MT system as is. Those corresponding to RQ2 were produced by our emotion-based procedure to recover the original emotion connotations, and those corresponding to RQ3 were selected by the same emotion-based procedure, when used to transfer emotions.

RQ	Emo.	Sentence
	G	*Feeling guilt after greed, buying chocolate and pigging out to the point of feeling sick, especially as I am fat.*
1	D	Feelings of greed, buying chocolate and exploitation to the point of nausea, mainly because I'm fat.
2	G	Feeling guilty about greed, buying chocolate and feeling sick, especially because I'm fat.
3	Sh	Feelings of greed, buying chocolate and feeling ill, mainly because I'm fat.
	F	*When I was first exposed to the dead bodies, for dissecting purposes at the school of medicine.*
1	D	When I was first confronted with the corpses to dissect them in medical school.
2	F	The first time I was confronted with the bodies, I dissected them in the medical school.
3	F	The first time I was confronted with the bodies, I dissected them in the medical school.
	Sa	*When my sister had the still born child, she was emotionally very deep down, and it took her a long time to recover.*
1	J	When my sister gave birth to the baby, she was very emotional and it took a long time for her to recover.
2	Sa	When my sister had the baby, she was emotionally very deep inside and it took a long time for her to recover.
3	A	When my sister had the baby, she was emotionally very low and it took a long time for her to recover.
	A	*During a recent meeting, Mr. A showed his excitement and overindulged in the notes delivered. Though his curiosity could not be blamed, his way of accquiring knowledge was an extreme behaviour e.g he always tried to know what I was reading and gained everything he could.*

1	D	During a recent meeting, Mr. A. showed his enthusiasm and left himself to the notes handed down. Although his curiosity could not be reproached, his way of acquiring knowledge was extreme, i.e. he always tried to know what I was reading and gained everything he could.
2	A	During one recent meeting, Mr. A. showed his enthusiasm and indulged excessively in the handed down notes. Although his curiosity could not be blamed, his way of acquiring knowledge was extreme, i.e. he always tried to know what I was reading and gained all he could.
3	Sa	During a recent meeting, Mr. A. showed his enthusiasm and revelled excessively in the notes handed down. Although he could not be blamed for his curiosity, his way of acquiring knowledge was extreme, that is, he always tried to know what I was reading and gained everything he could.
	D	*3 years ago I served in the army. Once a collegue denounced me because of a delict, which is usually committed. I was arrested for 3 days. I still detest this man.*
1	G	I served in the Army three years ago. A colleague once reported me for a crime that is normally committed. I was arrested for three days. I still loathe this man.
2	D	I served in the military three years ago. One time, a colleague reported me for a crime that is usually committed. I was arrested three days ago. I still detest that man.
3	Sa	Three years ago I was in the army. On one occasion a colleague reported me for an offence that is usually committed. I've been detained for three days. I still despair of this man.
	A	*When another fellow worker decided to leave the company. We had been very close and we would not be able to work with eachother any longer.*
1	Sa	When another employee decided to leave the company. We were very close and couldn't work together.
2	A	As another employee decided to leave the firm. We were close and couldn't work together any more.
3	G	When one more employee decided to leave the company. We were very close and could no longer work with one another.

Table B.12: Examples for the three research questions tackled with our backtranslation-based methodology on ISEAR. A: anger, D: disgust, F: fear, G: guilt, J: joy, Sa: sadness, Sh: shame.

4 Characterizing Emotions in Frame Semantics

This section provides details on the analyses of Chapter 6.

4.1 Corpus Labeling (Emotions)

All training parameters were kept constant for the 35 models obtained to find an automatic emotion annotator for COCA (see Page 194). The classifiers were fine-tuned for 10 epochs, setting a learning rate of $2*10^{-5}$ a dropout rate of 0.2, and a batch size of 32. We used AdamW as optimizer.

For a comparison to the BERT-based model selection, we experimented with a RoBERTa-based (Zhuang et al., 2021) emotion annotator trained on the whole concatenation of corpora (D) described on Page 194. While the latter yielded superior results when evaluated on the in-domain data, it deteriorated on the manually annotated sample of COCA as out-of-domain data. Results are reported in Table B.13.

Two viable alternatives for the automatic emotion annotation step could have been: (1) to use two classifiers, having high precision for either of the considered labels – i.e., one dedicated to the labeling of the emotional category and one for neutral category, which could arguably be more trustworthy, and (2) to accept texts as emotional or neutral if the probability with which the classifier assigns a label exceeds a given threshold. However, the first case would pose the problem of deciding how to treat texts for which the two models are in disagreement with one another. In the other case, we would lose substantial data. The decision to adopt an individual emotion labeler, with a reasonable F_1, bypasses both issues.

Adopting an annotation approach entirely based on human judgments would not be unproblematic either: sources compiled via crowdsourcing are noisy (Wauthier and Jordan, 2011); on the other hand, annotations conducted by expert coders cover smaller data, which hampers the attempt to draw empirical observations. We forgo the latter. Indeed, when it comes to judging emotions, the noisiness problem characterizes all human-based annotations, because the task is extremely subjective and therefore can lead to extreme disagreements, irrespective of how trained the coders are (cf. Chapter 4). Moreover, in the absence of the writers' annotations, no ultimate ground-truth holds with emotions, especially when assessing those that people are

	BERT-based	RoBERTa-based
D	F_1 =.83	F_1 = .86
COCA sample	F_1 =.69	F_1 = .55

Table B.13: BERT- and RoBERTa-based classifiers performance when trained and tested in domain (*D*) vs. trained on *D* and tested on the COCA sample with the majority vote treated as ground truth.

personally elicited by a text. Should the results of our analysis be due to systematic misclassifications of the automatic annotator, we could still assume that similar "errors" are to be found among humans.

4.2 Frames–Emotionality PMI Dictionary

Table B.14 reports the frame semantics dictionary built via the automatic annotation of COCA with emotionality and frames. The dictionary is sorted in descending order with respect to the first column (Emo. PMI), indicating the PMI value between a given frame and the label *emotional*. The right-most column (Neut. PMI) contains the values computed between a frame and the label *neutral*.

Frame	Emo. PMI	Neut. PMI
DYING	0.87775	–
EMOTIONS_BY_STIMULUS	0.8746	-7.71076
DISGRACEFUL_SITUATION	0.87124	-6.663
JUDGMENT_DIRECT_ADDRESS	0.86918	-6.26642
JUST_FOUND_OUT	0.86825	-6.11913
FEAR	0.86719	-5.9656
MAKING_FACES	0.85423	-4.81698
REASSURING	0.82983	-3.80252
EMOTION_ACTIVE	0.82699	-3.72068
FACIAL_EXPRESSION	0.8121	-3.357
CAUSE_EMOTION	0.79891	-3.09941
EXPERIENCER_OBJ	0.74992	-2.42637
REWARDS_AND_PUNISHMENTS	0.74886	-2.41497
EXPERIENCER_FOCUS	0.73668	-2.29067
FAIRNESS_EVALUATION	0.73531	-2.27741
CONTRITION	0.72503	-2.18195
SENTENCING	0.69978	-1.97355
CAUSE_TO_START	0.69752	-1.95651
COMMUNICATION_NOISE	0.68663	-1.87715
STIMULUS_FOCUS	0.68181	-1.84362

ACCURACY	0.6705	-1.76816
BUNGLING	0.66682	-1.74457
MENTAL_STIMULUS_STIMULUS_FOCUS	0.66157	-1.71167
LUCK	0.65953	-1.6991
MENTAL_PROPERTY	0.64811	-1.63107
PROTEST	0.64351	-1.6047
MAKE_NOISE	0.6357	-1.56118
BODY_MARK	0.62973	-1.52891
EMOTION_DIRECTED	0.62822	-1.5209
SATISFYING	0.61445	-1.4501
COGITATION	0.61442	-1.44993
ROTTING	0.6104	-1.43002
JUDGMENT	0.59331	-1.34891
SUCCESS_OR_FAILURE	0.59012	-1.33432
FEELING	0.58776	-1.3237
CHEMICAL-SENSE_DESCRIPTION	0.57666	-1.27486
DESIRABILITY	0.57653	-1.2743
FRUGALITY	0.57039	-1.24815
AGREE_OR_REFUSE_TO_ACT	0.56606	-1.23005
KILLING	0.56557	-1.22804
AESTHETICS	0.56478	-1.22475
BEAT_OPPONENT	0.56246	-1.21523
FIRING	0.56165	-1.21192
DESTROYING	0.55641	-1.1907
CHAOS	0.55605	-1.18926
TERRORISM	0.55436	-1.18255
DARING	0.55125	-1.1702
VERDICT	0.54022	-1.12758
FINISH_COMPETITION	0.53862	-1.12155
SOCIABILITY	0.53785	-1.11864
OFFENSES	0.53689	-1.11502
DESERVING	0.53569	-1.11055
RESPOND_TO_PROPOSAL	0.53254	-1.09883
CERTAINTY	0.52959	-1.08797
INSTITUTIONALIZATION	0.52529	-1.0723
DEATH	0.51176	-1.02445
RECOVERY	0.50716	-1.00864
SUASION	0.50364	-0.99669
OMEN	0.50199	-0.9911
RISKY_SITUATION	0.49942	-0.98252
KIDNAPPING	0.49442	-0.96595
GUILT_OR_INNOCENCE	0.49273	-0.96042
CAUSE_TO_EXPERIENCE	0.48958	-0.95018
SUBJECTIVE_INFLUENCE	0.48658	-0.94051
BEING_QUESTIONABLE	0.48493	-0.93522
PROMINENCE	0.48253	-0.92758
CAUSE_HARM	0.48216	-0.92638
REVENGE	0.47895	-0.91627

VOCALIZATIONS	0.47741	-0.91144
CATASTROPHE	0.47511	-0.90427
MISDEED	0.46958	-0.88724
ARREST	0.46761	-0.88123
PREVENT_OR_ALLOW_POSSESSION	0.46676	-0.87865
BIOLOGICAL_URGE	0.45938	-0.85647
GRASP	0.45383	-0.8401
IMPRISONMENT	0.45383	-0.8401
DIFFICULTY	0.45207	-0.83497
ARRAIGNMENT	0.44493	-0.81437
MORALITY_EVALUATION	0.44273	-0.80809
COMING_TO_BELIEVE	0.44084	-0.80274
STINGINESS	0.43585	-0.78871
ACCOMPLISHMENT	0.43559	-0.78798
VIOLENCE	0.43534	-0.78731
SOCIAL_INTERACTION_EVALUATION	0.43419	-0.78408
SUCCESSFUL_ACTION	0.43287	-0.78043
RENDER_NONFUNCTIONAL	0.42984	-0.77209
ARTIFICIALITY	0.42913	-0.77012
FLEEING	0.42261	-0.75242
UNEMPLOYMENT_RATE	0.41839	-0.74109
WARNING	0.41543	-0.73321
FORGING	0.41496	-0.73198
RENUNCIATION	0.41181	-0.72368
HIT_OR_MISS	0.40861	-0.71531
PREDICAMENT	0.4052	-0.70649
ASSISTANCE	0.39734	-0.68641
DESIRING	0.39515	-0.68086
IMPROVEMENT_OR_DECLINE	0.3938	-0.67748
WEALTHINESS	0.39301	-0.6755
CORRECTNESS	0.39288	-0.67517
COMMITMENT	0.39231	-0.67375
RITE	0.39206	-0.67312
ENTERING_OF_PLEA	0.38969	-0.66723
REBELLION	0.38535	-0.6565
LEVEL_OF_FORCE_EXERTION	0.38402	-0.65324
LINGUISTIC_MEANING	0.38113	-0.64617
COMPLAINING	0.37852	-0.63984
REASONING	0.37568	-0.633
ATTACK	0.37484	-0.63099
REPEL	0.37382	-0.62853
HOSTILE_ENCOUNTER	0.37301	-0.6266
PEOPLE_BY_MORALITY	0.37295	-0.62645
ENDANGERING	0.3709	-0.62158
CAUSE_TO_FRAGMENT	0.36873	-0.61644
DOMINATE_COMPETITOR	0.36816	-0.61508
SOCIAL_DESIRABILITY	0.3644	-0.60626
RESCUING	0.36327	-0.60362

BOARD_VEHICLE	0.36264	-0.60213
JUSTIFYING	0.36193	-0.60049
PREVARICATION	0.36104	-0.59842
JUDGMENT_COMMUNICATION	0.35932	-0.59444
WILLINGNESS	0.35895	-0.59358
SUBVERSION	0.35429	-0.58286
SENSATION	0.3523	-0.57833
COMPATIBILITY	0.35135	-0.57616
RESOLVE_PROBLEM	0.35005	-0.57321
EXPERIENCE_BODILY_HARM	0.34949	-0.57196
INCLINATION	0.34817	-0.56896
ARSON	0.34517	-0.56224
EXPRESSING_PUBLICLY	0.34278	-0.5569
MANIPULATE_INTO_DOING	0.34237	-0.55599
LAUNCH_PROCESS	0.33509	-0.53992
MILITARY_OPERATION	0.3346	-0.53884
MEDICAL_SPECIALTIES	0.33252	-0.53429
REFORMING_A_SYSTEM	0.3317	-0.53251
MEDICAL_CONDITIONS	0.33155	-0.5322
TRIGGERING	0.33088	-0.53074
ECONOMY	0.33054	-0.52999
TEMPERATURE	0.33009	-0.52901
CO-ASSOCIATION	0.32637	-0.521
EXAMINATION	0.32633	-0.52092
EXPECTATION	0.32604	-0.5203
EXPEND_RESOURCE	0.32583	-0.51985
JUDGMENT_OF_INTENSITY	0.32557	-0.51927
INFECTING	0.32085	-0.5092
AFFIRM_OR_DENY	0.31541	-0.49772
BEHIND_THE_SCENES	0.30794	-0.48218
APPELLATIONS	0.30742	-0.48111
TRUST	0.3072	-0.48065
RUN_RISK	0.30563	-0.47741
RIDE_VEHICLE	0.30412	-0.47434
EVENT	0.30296	-0.47196
OPPORTUNITY	0.30284	-0.47172
BEING_RELEVANT	0.30256	-0.47114
DEAD_OR_ALIVE	0.30023	-0.46639
IRREGULAR_COMBATANTS	0.29992	-0.46576
AWARENESS_STATUS	0.29776	-0.46139
DYNAMISM	0.29437	-0.45457
ENDEAVOR_FAILURE	0.29426	-0.45435
INVADING	0.29419	-0.4542
BEING_OPERATIONAL	0.29383	-0.45348
THEFT	0.29327	-0.45236
FUNDING	0.29278	-0.45138
DURATION_RELATION	0.29278	-0.45138
CHANGE_OF_LEADERSHIP	0.29266	-0.45113

FAME	0.29095	-0.44771
HOSPITALITY	0.28913	-0.4441
QUARRELING	0.28825	-0.44235
BEING_AT_RISK	0.28533	-0.43659
OPINION	0.28081	-0.42773
MEDICAL_INTERVENTION	0.2762	-0.41878
BEARING_ARMS	0.27459	-0.41566
REQUIRED_EVENT	0.27404	-0.41461
PEOPLE_BY_RELIGION	0.26914	-0.40522
REVEAL_SECRET	0.26466	-0.39673
MEDICAL_INTERACTION_SCENARIO	0.26456	-0.39654
PROLIFERATING_IN_NUMBER	0.26368	-0.39489
EDUCATION_TEACHING	0.26209	-0.39189
ESCAPING	0.26064	-0.38916
CAUSE_IMPACT	0.25551	-0.3796
CAUSE_TO_RESUME	0.25297	-0.37491
DAMAGING	0.2504	-0.37018
PRISON	0.24942	-0.36838
MAKE_COMPROMISE	0.24929	-0.36814
PRECARIOUSNESS	0.24876	-0.36717
MEET_SPECIFICATIONS	0.24795	-0.36569
TRIAL	0.24725	-0.36442
COMMITTING_CRIME	0.24701	-0.36398
MOTION_NOISE	0.24654	-0.36311
SURVIVING	0.2459	-0.36194
ATTEMPT	0.24477	-0.35989
MAKE_AGREEMENT_ON_ACTION	0.24402	-0.35853
SURRENDERING	0.24397	-0.35844
REPRESENTATIVE	0.24339	-0.35739
BREATHING	0.24301	-0.35671
EXECUTION	0.24223	-0.3553
CONFRONTING_PROBLEM	0.24132	-0.35365
EVENTIVE_AFFECTING	0.23958	-0.35052
TOURING	0.23943	-0.35026
ATTITUDE_DESCRIPTION	0.23906	-0.34959
CURE	0.23823	-0.34811
EXPERTISE	0.23464	-0.34168
LEGAL_RULINGS	0.23427	-0.34103
AWARENESS	0.23426	-0.341
ABANDONMENT	0.23379	-0.34017
FUGITIVE	0.23057	-0.33447
THRIVING	0.23024	-0.33388
CAUSE_CHANGE_OF_STRENGTH	0.22997	-0.33341
RESPONSE	0.2286	-0.331
COMPETITION	0.22808	-0.33007
UNDERGOING	0.22742	-0.32891
CONFERRING_BENEFIT	0.22708	-0.32832
SHOOT_PROJECTILES	0.22705	-0.32827

PROGRESSION	0.22534	-0.32527
COGNITIVE_CONNECTION	0.22411	-0.32313
REMEMBERING_INFORMATION	0.22239	-0.32014
CAUSE_TO_WAKE	0.221	-0.31772
BEING_IN_EFFECT	0.21937	-0.3149
PROCESS_START	0.2193	-0.31477
EMERGENCY_FIRE	0.2191	-0.31443
INTENTIONALLY_ACT	0.21788	-0.31231
LAW_ENFORCEMENT_AGENCY	0.21752	-0.3117
VEHICLE_SUBPART	0.21223	-0.30263
WORK	0.20934	-0.29771
ACTIVITY_PREPARE	0.20846	-0.29622
TAKING	0.20841	-0.29613
NEEDING	0.20528	-0.29084
READING_PERCEPTION	0.20519	-0.2907
DECIDING	0.20432	-0.28924
REJUVENATION	0.20231	-0.28586
PROPER_REFERENCE	0.19998	-0.28196
PEOPLE	0.19889	-0.28015
CAUSE_TO_MAKE_PROGRESS	0.19799	-0.27865
HIT_TARGET	0.19768	-0.27814
PROBABILITY	0.19525	-0.27412
CONQUERING	0.19177	-0.26837
ATTENDING	0.19155	-0.26802
SUSPICION	0.18918	-0.26413
DESTINY	0.18834	-0.26277
PEOPLE_BY_ORIGIN	0.18728	-0.26104
TOXIC_SUBSTANCE	0.18268	-0.25357
BODY_MOVEMENT	0.18231	-0.25297
MEDICAL_PROFESSIONALS	0.18171	-0.252
EXCRETING	0.1803	-0.24974
BECOMING	0.18004	-0.24931
DISCUSSION	0.1789	-0.24749
TEMPORARY_LEAVE	0.17775	-0.24563
ALLIANCE	0.17716	-0.24469
EXTRADITION	0.17604	-0.24291
CHANGE_EVENT_TIME	0.17558	-0.24217
KINSHIP	0.17556	-0.24214
CHATTING	0.17551	-0.24207
JURY_DELIBERATION	0.17301	-0.23809
POSING_AS	0.17117	-0.23516
PARTICIPATION	0.17011	-0.2335
USEFULNESS	0.17005	-0.2334
TRAVERSING	0.16744	-0.22928
OBJECTIVE_INFLUENCE	0.16599	-0.22701
PROCESS_END	0.16563	-0.22645
HERALDING	0.16539	-0.22607
SEEKING_TO_ACHIEVE	0.16524	-0.22583

CAUSATION	0.16434	-0.22443
DEFENDING	0.16355	-0.2232
PERFORMING_ARTS	0.16344	-0.22302
SURRENDERING_POSSESSION	0.16272	-0.2219
CONDUCT	0.16187	-0.22057
RECEIVING	0.16121	-0.21954
GETTING	0.15979	-0.21735
PREDICTING	0.15945	-0.21681
OPERATIONAL_TESTING	0.15798	-0.21453
CAUSE_TO_MOVE_IN_PLACE	0.15743	-0.21369
ATTEMPT_MEANS	0.15629	-0.21193
SOCIAL_EVENT	0.15618	-0.21176
IMPACT	0.15605	-0.21156
LOCATING	0.15555	-0.21079
PROCESS_CONTINUE	0.15357	-0.20774
CRIMINAL_INVESTIGATION	0.14551	-0.19546
UNDERGO_CHANGE	0.14448	-0.19392
PEOPLE_BY_AGE	0.14222	-0.19051
ACTIVITY_RESUME	0.1419	-0.19003
ACTIVITY_READY_STATE	0.13962	-0.18661
SELF_CONTROL	0.13958	-0.18655
REGARD	0.13761	-0.18361
ACHIEVING_FIRST	0.13515	-0.17994
FEIGNING	0.13428	-0.17865
BEING_EMPLOYED	0.133	-0.17676
FIGHTING_ACTIVITY	0.13165	-0.17476
TRAVEL	0.1316	-0.17469
LEVEL_OF_FORCE_RESISTANCE	0.13075	-0.17343
INFRASTRUCTURE	0.13053	-0.17311
MEMORY	0.12898	-0.17083
FORGOING	0.12567	-0.16598
HISTORY	0.1245	-0.16427
RELATION	0.12401	-0.16355
POLITICAL_LOCALES	0.12397	-0.16349
RESEARCH	0.12346	-0.16276
AVOIDING	0.12331	-0.16253
CALENDRIC_UNIT	0.122	-0.16062
MILITARY	0.11884	-0.15605
READING_ACTIVITY	0.1172	-0.15368
TRENDINESS	0.11571	-0.15154
MEMBER_OF_MILITARY	0.11547	-0.15119
SETTING_FIRE	0.1149	-0.15038
SUPPORTING	0.11368	-0.14863
FOREIGN_OR_DOMESTIC_COUNTRY	0.11352	-0.1484
BEING_OBLIGATED	0.11268	-0.1472
IDIOSYNCRASY	0.11261	-0.14709
ATTENTION	0.11101	-0.14482
ASSESSING	0.10755	-0.13989

LEGALITY	0.10752	-0.13986
PROJECT	0.10673	-0.13874
PERSONAL_RELATIONSHIP	0.10652	-0.13843
AGGREGATE	0.10376	-0.13454
BEING_LOCATED	0.10225	-0.13242
ATTEMPT_SUASION	0.10101	-0.13067
DOMINATE_SITUATION	0.10097	-0.13062
TYPE	0.10035	-0.12976
SHOPPING	0.09659	-0.12449
COMMUNICATION_RESPONSE	0.09446	-0.12154
TELLING	0.09362	-0.12038
LIGHT_MOVEMENT	0.09362	-0.12037
EXPERIMENTATION	0.09267	-0.11906
CHEMICAL_POTENCY	0.09066	-0.11629
EVOKING	0.08968	-0.11494
NONCOMBATANT	0.08925	-0.11435
QUITTING_A_PLACE	0.08902	-0.11403
PEOPLE_BY_JURISDICTION	0.08869	-0.11357
POINT_OF_DISPUTE	0.08788	-0.11246
STRICTNESS	0.08738	-0.11178
PROCESS_STOP	0.08668	-0.11082
TIMETABLE	0.08395	-0.10709
MOVING_IN_PLACE	0.08345	-0.10641
CAUSE_CHANGE_OF_PHASE	0.08333	-0.10625
PROTECTING	0.08255	-0.10518
REASON	0.08233	-0.10489
MAKE_ACQUAINTANCE	0.08007	-0.10182
SUITABILITY	0.07907	-0.10046
GIVE_IMPRESSION	0.07739	-0.09819
UNATTRIBUTED_INFORMATION	0.07686	-0.09749
IMPORTANCE	0.07683	-0.09745
PERCEPTION_EXPERIENCE	0.07639	-0.09685
CONTINUED_STATE_OF_AFFAIRS	0.07605	-0.09639
ACTIVITY_STOP	0.07563	-0.09583
CREATE_PHYSICAL_ARTWORK	0.07418	-0.09388
LIVELY_PLACE	0.07305	-0.09237
RASHNESS	0.07172	-0.09059
LEADERSHIP	0.07077	-0.08932
BEING_IN_CAPTIVITY	0.07037	-0.08878
SOUNDS	0.06805	-0.0857
OBVIOUSNESS	0.06745	-0.0849
ACTIVITY_FINISH	0.06745	-0.0849
FIRE_BURNING	0.06669	-0.08389
BREAKING_OUT_CAPTIVE	0.06647	-0.0836
ORGANIZATION	0.06413	-0.0805
PROHIBITING_OR_LICENSING	0.06338	-0.07952
LOCALE_BY_EVENT	0.0624	-0.07823
IMPORT_EXPORT_SCENARIO	0.06196	-0.07764

OPERATE_VEHICLE	0.06187	-0.07752
SIMULTANEITY	0.06173	-0.07734
FREQUENCY	0.06028	-0.07543
INTOXICANTS	0.0596	-0.07455
TRAP	0.05955	-0.07448
SUFFICIENCY	0.05936	-0.07424
SPEED_DESCRIPTION	0.05866	-0.07332
SUBORDINATES_AND_SUPERIORS	0.05849	-0.07309
RESIDENCE	0.05672	-0.07078
INTERRUPT_PROCESS	0.05641	-0.07037
ACCOUTREMENTS	0.05629	-0.07022
PIRACY	0.0559	-0.06972
PRACTICE	0.05424	-0.06755
SIMILARITY	0.05414	-0.06743
GIVING	0.05253	-0.06533
BEING_ATTACHED	0.05165	-0.06419
PROCESS	0.05101	-0.06337
COMMUNICATION	0.05094	-0.06328
CHOOSING	0.05007	-0.06215
BECOMING_A_MEMBER	0.04972	-0.0617
EMPHASIZING	0.04945	-0.06135
CAUSE_MOTION	0.04929	-0.06115
BESIEGING	0.0452	-0.0559
EXECUTE_PLAN	0.04496	-0.05559
FORMING_RELATIONSHIPS	0.04365	-0.05391
HINDERING	0.04332	-0.05348
INSTITUTIONS	0.04274	-0.05275
PEOPLE_BY_VOCATION	0.04154	-0.05121
QUANTITY	0.04003	-0.04929
ACTIVITY_START	0.03853	-0.04739
STATEMENT	0.03799	-0.04671
PEOPLE_BY_RESIDENCE	0.03697	-0.04542
RELEASING	0.03622	-0.04447
FIELDS	0.03591	-0.04408
USED_UP	0.03546	-0.04351
GRAPH_SHAPE	0.03485	-0.04274
RATE_DESCRIPTION	0.03319	-0.04065
MEMBERSHIP	0.03254	-0.03984
ORDINAL_NUMBERS	0.03163	-0.0387
SCHEDULING	0.02975	-0.03634
AGE	0.028	-0.03416
HEARSAY	0.02755	-0.0336
ENFORCING	0.02592	-0.03157
INDIGENOUS_ORIGIN	0.02323	-0.02824
MEASURE_BY_ACTION	0.02319	-0.02818
PERCEPTION_ACTIVE	0.02151	-0.02611
POSSESSION	0.01765	-0.02137
EVIDENCE	0.01735	-0.02099

TEMPORAL_SUBREGION	0.01574	-0.01903
SOCIAL_EVENT_COLLECTIVE	0.0143	-0.01726
BECOMING_AWARE	0.01387	-0.01674
PURPOSE	0.01335	-0.01611
HISTORIC_EVENT	0.01276	-0.01539
ERASING	0.01268	-0.01528
RANKED_EXPECTATION	0.01236	-0.0149
BIOLOGICAL_AREA	0.01102	-0.01326
ORIGIN	0.01053	-0.01268
SYSTEM_COMPLEXITY	0.01036	-0.01246
NOTIFICATION_OF_CHARGES	0.00801	-0.00962
HAVE_AS_REQUIREMENT	0.00792	-0.00952
ACTIVITY_PAUSE	0.00773	-0.00928
STATE_OF_ENTITY	0.00753	-0.00904
LOCALE_BY_USE	0.00659	-0.0079
COMPLETENESS	0.00654	-0.00785
ROBBERY	0.00513	-0.00614
TEMPORAL_COLLOCATION	0.00499	-0.00598
WORD_RELATIONS	0.00489	-0.00586
LIKELIHOOD	0.00478	-0.00573
INGESTION	0.00462	-0.00554
EMPLOYING	0.00409	-0.00489
PRESENCE	0.00262	-0.00313
CHANGE_RESISTANCE	-1.45615	0.81463
EXEMPLARINESS	-1.35255	0.78781
PLANTS	-1.25976	0.76161
LOCALE_BY_OWNERSHIP	-1.20972	0.74656
PATROLLING	-1.17288	0.73503
DELIMITATION_OF_DIVERSITY	-1.02236	0.68365
PATH_TRAVELED	-0.93252	0.64932
SUBSTANCE_BY_PHASE	-0.91516	0.64233
MOTION_DIRECTIONAL	-0.87157	0.62427
CONTROL	-0.82781	0.60533
ARMOR	-0.81118	0.59792
SHARING	-0.79211	0.58926
DIRECTIONAL_LOCATIVE_RELATION	-0.78896	0.58782
INTENTIONAL_TRAVERSING	-0.77699	0.58228
MEASURE_AREA	-0.75566	0.57224
STORING	-0.74854	0.56885
MARGIN_OF_RESOLUTION	-0.74802	0.56859
RESERVING	-0.73224	0.56097
BIOLOGICAL_CLASSIFICATION	-0.72054	0.55524
ESTIMATED_VALUE	-0.69164	0.54076
CONTAINING	-0.68288	0.5363
COMMERCIAL_TRANSACTION	-0.6662	0.52766
BEING_WET	-0.66138	0.52514
MEASURABLE_ATTRIBUTES	-0.65752	0.52311
CAPACITY	-0.65083	0.51957

BECOMING_ATTACHED	-0.64729	0.51769
ROPE_MANIPULATION	-0.63156	0.50925
BECOMING_DRY	-0.62632	0.5064
USING_RESOURCE	-0.61868	0.50223
SCOURING	-0.59403	0.48851
MEASURE_LINEAR_EXTENT	-0.58734	0.48472
CLOTHING_PARTS	-0.57823	0.47953
MEASURE_VOLUME	-0.57468	0.47748
ADJACENCY	-0.56293	0.47068
COLONIZATION	-0.56292	0.47067
RATE_QUANTIFICATION	-0.56283	0.47062
CHANGE_POST-STATE	-0.56008	0.46901
PATTERN	-0.54428	0.45969
HEALTH_RESPONSE	-0.52927	0.45068
LEFT_TO_DO	-0.52101	0.44565
DISPERSAL	-0.52054	0.44536
NON-GRADABLE_PROXIMITY	-0.51868	0.44423
GROOMING	-0.5122	0.44024
BILLING	-0.50988	0.43881
RECORDS	-0.50152	0.43361
CAUSE_FLUIDIC_MOTION	-0.49246	0.42792
DISTRIBUTED_POSITION	-0.48926	0.42591
OPTICAL_IMAGE	-0.48215	0.42138
TERMS_OF_AGREEMENT	-0.47946	0.41966
BECOMING_SILENT	-0.47681	0.41797
RELATIONAL_NATURAL_FEATURES	-0.47558	0.41718
ESTIMATING	-0.47431	0.41635
REFERRING_BY_NAME	-0.46718	0.41174
LOCALE_BY_CHARACTERISTIC_ENTITY	-0.46239	0.40863
CAUSE_TEMPERATURE_CHANGE	-0.45759	0.40548
CONTAINERS	-0.45677	0.40495
INHIBIT_MOVEMENT	-0.45653	0.40479
LABOR_PRODUCT	-0.44999	0.40048
GESTURE	-0.44626	0.398
SURROUNDING	-0.44554	0.39752
ENTITY	-0.43831	0.39269
BODY_DESCRIPTION_HOLISTIC	-0.43321	0.38925
CORPORAL_PUNISHMENT	-0.43077	0.38761
GATHERING_UP	-0.43068	0.38754
REMAINDER	-0.43058	0.38748
SUMMARIZING	-0.42694	0.385
SEPARATING	-0.42688	0.38497
LIMITATION	-0.42197	0.38162
INSTALLING	-0.42014	0.38036
SHAPES	-0.41882	0.37946
INGREDIENTS	-0.41526	0.377
SEX	-0.41183	0.37464
CONNECTORS	-0.40398	0.36918

CHANGE_EVENT_DURATION	-0.40012	0.36647
GUSTO	-0.39831	0.36521
FORGIVENESS	-0.39731	0.3645
MEASURE_DURATION	-0.39447	0.36249
LENDING	-0.39282	0.36133
SCOPE	-0.39272	0.36126
NATURAL_FEATURES	-0.39094	0.35999
VOLUBILITY	-0.38455	0.35545
CHANGE_TOOL	-0.38227	0.35381
CONNECTING_ARCHITECTURE	-0.37677	0.34986
EXPLAINING_THE_FACTS	-0.37364	0.34759
DIRECTION	-0.37056	0.34536
SUBSTANCE	-0.36807	0.34355
INSPECTING	-0.36764	0.34323
DUPLICATION	-0.36265	0.33958
EXTREME_VALUE	-0.35959	0.33734
HUNTING	-0.35948	0.33725
PART_INNER_OUTER	-0.35882	0.33677
HAVING_OR_LACKING_ACCESS	-0.35871	0.33669
SPATIAL_CONTACT	-0.35728	0.33563
ARCHITECTURAL_PART	-0.35637	0.33496
APPLY_HEAT	-0.3536	0.33291
FOOD_GATHERING	-0.35228	0.33192
BEING_NAMED	-0.35195	0.33168
PARTITIVE	-0.35166	0.33147
DIMENSION	-0.34728	0.3282
GRADABLE_PROXIMITY	-0.34465	0.32623
PEOPLE_ALONG_POLITICAL_SPECTRUM	-0.34088	0.32339
LIMITING	-0.33872	0.32177
IMPOSING_OBLIGATION	-0.33865	0.32171
APPOINTING	-0.33857	0.32165
EXEMPLAR	-0.33701	0.32047
PART_ORIENTATIONAL	-0.32992	0.31509
LOCATION_OF_LIGHT	-0.32871	0.31415
LABELING	-0.32568	0.31184
AMALGAMATION	-0.32531	0.31155
CAUSE_TO_AMALGAMATE	-0.3231	0.30985
MASS_MOTION	-0.32296	0.30975
IMPORTING	-0.31943	0.30702
STORE	-0.31516	0.30371
GRINDING	-0.31294	0.30198
SIZE	-0.31266	0.30176
ADDUCING	-0.31097	0.30044
BE_IN_AGREEMENT_ON_ACTION	-0.30929	0.29913
GOAL	-0.30894	0.29886
BECOMING_SEPARATED	-0.30888	0.29881
SOAKING_UP	-0.30857	0.29856
TRANSITION_TO_A_QUALITY	-0.3081	0.2982

CONTINGENCY	-0.30696	0.2973
BEING_IN_OPERATION	-0.30548	0.29613
MEASURE_MASS	-0.29725	0.28962
WAGERING	-0.29696	0.2894
GIZMO	-0.29678	0.28925
CHANGE_OF_TEMPERATURE	-0.2959	0.28856
ENCODING	-0.29453	0.28746
OBSCURITY	-0.29309	0.28631
BIOLOGICAL_ENTITY	-0.29118	0.28478
REPLACING	-0.29012	0.28393
BE_SUBSET_OF	-0.28765	0.28195
BEING_IN_CATEGORY	-0.28595	0.28058
FRONT_FOR	-0.28421	0.27918
TRANSFER	-0.28345	0.27857
PLACING	-0.27736	0.27362
HEDGING	-0.27647	0.27289
RATIFICATION	-0.27481	0.27154
COLOR	-0.27303	0.27008
DIVERSITY	-0.27089	0.26833
ECLIPSE	-0.27079	0.26825
NEGATIVE_CONDITIONAL	-0.27008	0.26766
CARRY_GOODS	-0.27006	0.26764
MEDICAL_INSTRUMENTS	-0.26935	0.26706
HAIR_CONFIGURATION	-0.26466	0.26319
CLOSURE	-0.26154	0.26059
FIRST_EXPERIENCE	-0.26065	0.25986
COMMUTATIVE_PROCESS	-0.25976	0.25911
FULLNESS	-0.25912	0.25858
AMASSING	-0.25859	0.25814
COMMERCE_SCENARIO	-0.25831	0.25791
EMERGENCY	-0.25793	0.25759
ATTACHING	-0.25737	0.25712
RETAINING	-0.2573	0.25706
BOUNDARY	-0.25677	0.25662
CARDINAL_NUMBERS	-0.25292	0.25339
SOLE_INSTANCE	-0.25234	0.2529
COMMERCE_SELL	-0.25198	0.2526
CHANGE_OF_PHASE	-0.25178	0.25243
SENT_ITEMS	-0.24817	0.24939
USING	-0.2478	0.24907
CAUSE_TO_END	-0.24692	0.24833
BAIL_DECISION	-0.24478	0.24651
PROPORTIONAL_QUANTITY	-0.24385	0.24572
INCLUSION	-0.24204	0.24418
SOURCE_OF_GETTING	-0.24172	0.24391
SUBMITTING_DOCUMENTS	-0.23965	0.24214
BODY_DECORATION	-0.23773	0.2405
REDIRECTING	-0.23411	0.23739

PROPORTION	-0.23199	0.23555
ACTUALLY_OCCURRING_ENTITY	-0.23143	0.23507
EVENT_INSTANCE	-0.23065	0.2344
RANK	-0.22871	0.23272
BUILDING	-0.22851	0.23254
INTERIOR_PROFILE_RELATION	-0.22416	0.22876
FLUIDIC_MOTION	-0.22361	0.22828
MAKING_ARRANGEMENTS	-0.22243	0.22725
REMOVING	-0.22118	0.22616
DISTINCTIVENESS	-0.22089	0.22591
COMING_UP_WITH	-0.21902	0.22426
PARTIALITY	-0.21747	0.2229
PERFORMERS	-0.21741	0.22285
EMPTYING	-0.21568	0.22133
WEARING	-0.21461	0.22038
CLOTHING	-0.21375	0.21962
MAKE_COGNITIVE_CONNECTION	-0.21249	0.2185
COMMERCE_PAY	-0.21211	0.21817
INFORMATION	-0.21105	0.21723
PRECIPITATION	-0.21059	0.21682
CAUSE_TO_MAKE_NOISE	-0.20988	0.21619
COMMONALITY	-0.20632	0.21302
EXCHANGE	-0.20525	0.21206
EXPANSION	-0.20497	0.21181
ACTIVITY_DONE_STATE	-0.20484	0.21169
MANIPULATION	-0.20465	0.21152
TEMPORARY_STAY	-0.20232	0.20943
PART_PIECE	-0.20166	0.20884
TYPICALITY	-0.20029	0.20762
PUTTING_OUT_FIRE	-0.19843	0.20593
SOCIAL_CONNECTION	-0.19809	0.20563
ROADWAYS	-0.19779	0.20535
NAME_CONFERRAL	-0.19687	0.20452
PRESENTATION_OF_MITIGATION	-0.19601	0.20375
FINING	-0.19501	0.20284
ISOLATED_PLACES	-0.19428	0.20218
COMMUNICATE_CATEGORIZATION	-0.18988	0.19818
FOOD	-0.18881	0.1972
RANGE	-0.18398	0.19277
POSTURE	-0.18389	0.1927
AMBIENT_TEMPERATURE	-0.18321	0.19206
OPENNESS	-0.18313	0.192
CUTTING	-0.18268	0.19158
BEING_ACTIVE	-0.18214	0.19108
AMOUNTING_TO	-0.18138	0.19038
SUPPLY	-0.17889	0.18808
STATE_CONTINUE	-0.17859	0.1878
BEING_BORN	-0.17759	0.18687

BUILDING_SUBPARTS	-0.16973	0.17955
DEGREE_OF_PROCESSING	-0.16846	0.17835
ADOPT_SELECTION	-0.16745	0.1774
FILLING	-0.16616	0.17619
AGRICULTURE	-0.16311	0.17331
OCCUPY_RANK	-0.16162	0.1719
LOCALE	-0.16083	0.17115
PREVENTING_OR_LETTING	-0.15825	0.1687
EVALUATIVE_COMPARISON	-0.15801	0.16847
PUBLISHING	-0.15795	0.16841
TAKE_PLACE_OF	-0.15414	0.16478
BEING_NECESSARY	-0.15028	0.16108
RELATIONAL_POLITICAL_LOCALES	-0.14972	0.16053
CREATING	-0.14952	0.16034
PRELIMINARIES	-0.14916	0.15999
EARNINGS_AND_LOSSES	-0.14836	0.15923
DETAINING	-0.14833	0.15919
DOCUMENTS	-0.14685	0.15777
AIMING	-0.14588	0.15683
NUCLEAR_PROCESS	-0.1452	0.15617
BODY_PARTS	-0.14501	0.15598
THERMODYNAMIC_PHASE	-0.14445	0.15544
INDIVIDUAL_HISTORY	-0.14419	0.15519
IMPRESSION	-0.14404	0.15505
NON-COMMUTATIVE_PROCESS	-0.13851	0.14965
LOCATIVE_RELATION	-0.13827	0.14942
ACTIVITY_ONGOING	-0.13734	0.14851
INCREMENT	-0.13652	0.14771
NETWORK	-0.13643	0.14762
IDENTICALITY	-0.13612	0.14731
CEASING_TO_BE	-0.1357	0.1469
DRESSING	-0.13426	0.14549
SCARCITY	-0.13379	0.14502
REST	-0.13352	0.14476
INSTANCE	-0.13295	0.1442
MANUFACTURING	-0.13266	0.14392
PUNCTUAL_PERCEPTION	-0.13195	0.14321
GROUND_UP	-0.13186	0.14312
SOUND_MOVEMENT	-0.13107	0.14235
COINCIDENCE	-0.13083	0.1421
NON-COMMUTATIVE_STATEMENT	-0.12919	0.14048
CAUSE_CHANGE	-0.12754	0.13885
INGEST_SUBSTANCE	-0.12741	0.13872
PHYSICAL_ARTWORKS	-0.12738	0.13869
FALL_ASLEEP	-0.12616	0.13747
EXPENSIVENESS	-0.12534	0.13666
COME_DOWN_WITH	-0.12444	0.13577
PART_WHOLE	-0.12441	0.13574

ACTIVE_SUBSTANCE	-0.12428	0.1356
CONCESSIVE	-0.12297	0.13429
MANNER_OF_LIFE	-0.12225	0.13358
FRIENDLY_OR_HOSTILE	-0.12225	0.13358
COLOR_QUALITIES	-0.12225	0.13358
HIDING_OBJECTS	-0.11996	0.13128
CRAFT	-0.11854	0.12986
EXPORTING	-0.1182	0.12952
AMMUNITION	-0.1166	0.1279
TAKING_TIME	-0.11637	0.12768
TRANSLATING	-0.11545	0.12675
ABOUNDING_WITH	-0.11538	0.12667
DENY_OR_GRANT_PERMISSION	-0.11459	0.12587
COLLABORATION	-0.11285	0.12411
QUANTIFIED_MASS	-0.1128	0.12407
BRINGING	-0.11169	0.12293
DOMAIN	-0.11035	0.12158
TOPIC	-0.10721	0.11838
SIGN_AGREEMENT	-0.1061	0.11725
ARRIVING	-0.10529	0.11642
SOUND_LEVEL	-0.10501	0.11612
SIMPLE_NAME	-0.10481	0.11593
ARTIFACT	-0.10183	0.11287
ARRANGING	-0.1	0.11098
PUBLIC_SERVICES	-0.09859	0.10952
RECORDING	-0.09811	0.10903
PRESERVING	-0.09782	0.10872
CHANGE_POSITION_ON_A_SCALE	-0.09724	0.10813
UNDERGO_TRANSFORMATION	-0.09707	0.10794
ALTERNATIVES	-0.09621	0.10705
HIRING	-0.0958	0.10663
PROCESS_COMPLETED_STATE	-0.09461	0.1054
COMMUTATIVE_STATEMENT	-0.09442	0.1052
MINING	-0.0928	0.10351
OFFERING	-0.09026	0.10086
POSITION_ON_A_SCALE	-0.08967	0.10024
RELATIVE_TIME	-0.08889	0.09943
COMMERCE_BUY	-0.08861	0.09913
BEING_OBLIGATORY	-0.08551	0.09588
PROCESSING_MATERIALS	-0.08509	0.09544
PATH_SHAPE	-0.08429	0.09459
TEAM	-0.08223	0.09241
SEQUENCE	-0.08085	0.09095
SENDING	-0.08034	0.09041
BUSINESSES	-0.07907	0.08906
COMING_TO_BE	-0.07787	0.08779
DESIRABLE_EVENT	-0.07626	0.08607
GIVING_IN	-0.07543	0.08519

MEANS	-0.07414	0.08381
SLEEP	-0.07342	0.08304
CAPITAL_STOCK	-0.0734	0.08301
WEATHER	-0.07226	0.08179
SYSTEM	-0.07124	0.0807
VERSION_SEQUENCE	-0.06931	0.07862
CANDIDNESS	-0.06681	0.07592
SIGN	-0.06641	0.07549
LAW	-0.0663	0.07537
PERFORMERS_AND_ROLES	-0.0652	0.07418
RESHAPING	-0.06496	0.07391
THWARTING	-0.06439	0.0733
CUSTOM	-0.06418	0.07307
OPERATING_A_SYSTEM	-0.06403	0.07291
REPORTING	-0.06336	0.07218
CATEGORIZATION	-0.06318	0.07198
REPRESENTING	-0.06213	0.07085
DIFFERENTIATION	-0.06195	0.07064
VERIFICATION	-0.06188	0.07057
COME_TOGETHER	-0.05844	0.06681
VEHICLE	-0.05477	0.06279
NOISE_MAKERS	-0.05193	0.05965
SURPASSING	-0.05164	0.05934
MONEY	-0.05093	0.05855
STAGE_OF_PROGRESS	-0.04857	0.05593
QUESTIONING	-0.04751	0.05475
CAPABILITY	-0.04609	0.05317
SHARPNESS	-0.04587	0.05292
OFFSHOOT	-0.04547	0.05247
CAUSE_CHANGE_OF_POSITION_ON_A_SCALE	-0.04452	0.05142
NEGATION	-0.04294	0.04966
DURATION_DESCRIPTION	-0.04084	0.04729
TAKING_SIDES	-0.03941	0.04569
CAUSE_CHANGE_OF_CONSISTENCY	-0.03909	0.04533
TEXT	-0.03889	0.04511
CONTACTING	-0.03766	0.04371
WAITING	-0.03661	0.04253
POSSIBILITY	-0.03619	0.04205
RELIANCE	-0.03564	0.04143
INTENTIONALLY_CREATE	-0.03545	0.04122
CAUSE_TO_PERCEIVE	-0.03435	0.03998
MOTION	-0.03418	0.03977
BUILDINGS	-0.03261	0.03799
SMUGGLING	-0.03235	0.0377
FAMILIARITY	-0.02996	0.03498
TEMPORAL_PATTERN	-0.02913	0.03403
SCRUTINY	-0.02867	0.0335
REQUEST	-0.02865	0.03348

PART_ORDERED_SEGMENTS	-0.0282	0.03296
COMMUNICATION_MANNER	-0.02735	0.032
WEAPON	-0.02694	0.03152
SECRECY_STATUS	-0.02607	0.03052
ELECTRICITY	-0.02286	0.02683
PREFERENCE	-0.02169	0.02548
EXISTENCE	-0.02053	0.02414
CAUSE_EXPANSION	-0.01926	0.02267
BEING_IN_CONTROL	-0.01904	0.02241
ATTRIBUTED_INFORMATION	-0.0182	0.02143
CATCHING_FIRE	-0.01733	0.02042
SELF_MOTION	-0.01716	0.02023
ANIMALS	-0.01707	0.02012
ACCOMPANIMENT	-0.01494	0.01764
FIRST_RANK	-0.01469	0.01735
DEPARTING	-0.01449	0.01712
COTHEME	-0.01436	0.01696
SEEKING	-0.01283	0.01517
COMPLIANCE	-0.01235	0.01461
DELIVERY	-0.011	0.01303
DOWNING	-0.0086	0.0102
JUDICIAL_BODY	-0.00838	0.00994
TEXT_CREATION	-0.00751	0.00891
ABUNDANCE	-0.00478	0.00569
VISITING	-0.00262	0.00312
TIME_VECTOR	-0.0015	0.00178
DEGREE	-0.00136	0.00162
INEFFABILITY	-0.00072	0.00085
SPECIFIC_INDIVIDUAL	–	1.13358
ATTENTION_GETTING	–	1.13358
ADDICTION	–	1.13358

Table B.14: Frames found in the COCA texts and their corresponding emotionality meanings, quantified as a PMI value between each frame and the emotion label *emotional* (Emo. PMI) and *neutral* (Neut. PMI).

4.3 Emotions–Frames Associations in crowd-enVENT

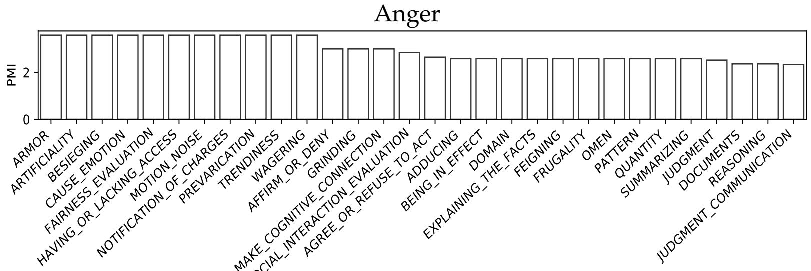
Using the same procedure applied on COCA, we computed the emotion–frame associations on crowd-enVENT: we estimate PMI for each pair (f, e) consisting of a frame (found with the SRL system used on COCA) and an emotion label (the categories prompting the event descriptions). Figure B.3 reports the top 30 frames for each emotion class, including the class *noemotion*. They provide a good semantic synthesis of the corpus. Some examples are:

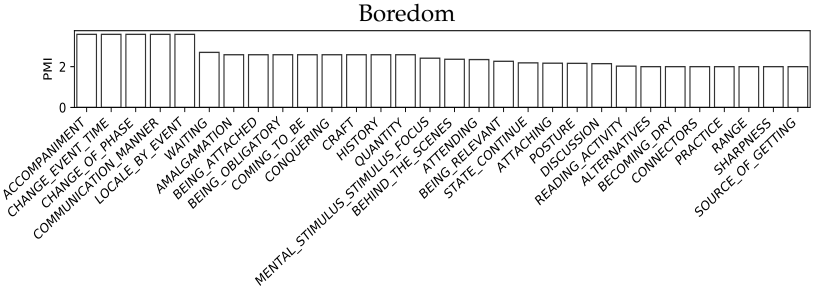
- PREVARICATION and FAIRNESS_EVALUATION highlighting the link between *anger* and an event perceived as unjust;
- GUILT_OR_INNOCENCE and INGESTION, which exemplify how *disgust* is talked about as a physical or a moral reaction;
- AWARENESS_STATUS and INVADING illustrating a component concerning the experiencer and a feature of events that accompany *fear*;
- BEAT_OPPONENT, with a prototypical connotation of *joy*, and ACCOMPLISHMENT, concerning *pride*;
- *sadness* is often instantiated with events of DEATH, KIDNAPPING, while *shame* with situations involving THEFT and ARREST;
- *surprise* evokes a notion of EXPECTATION, and descriptions labeled with *trust* evoke the frames SUPPORTING, RELIANCE, BEING_AT_RISK;
- *boredom* is strongly associated to WAITING and ATTENDING and the conceptually close *noemotion* class is associated to ordinary activities (e.g., FOOD, SHOPPING).

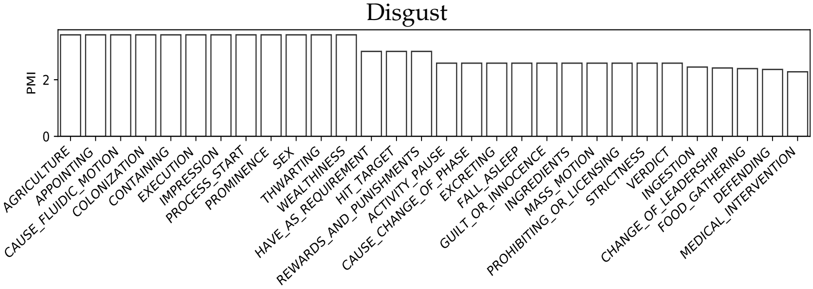
  The multiple descriptions regarding "exceptionally non-emotional" events in crowd-enVENT reflect on the frames most strongly associated to *noemotion*, some of which (e.g., PRISON) were associated to emotionality in COCA.

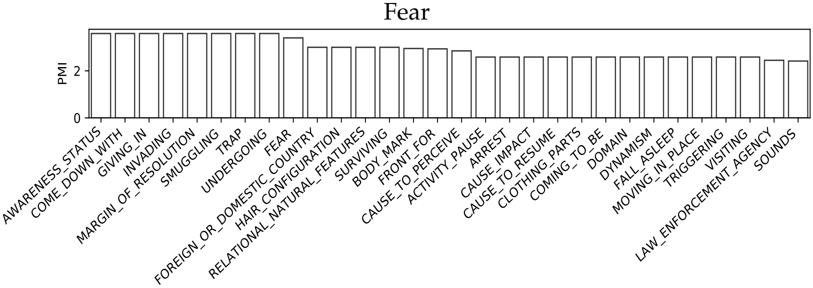
Note that only for some emotion classes FrameNet has a corresponding frame (e.g., FEAR, TRUST).

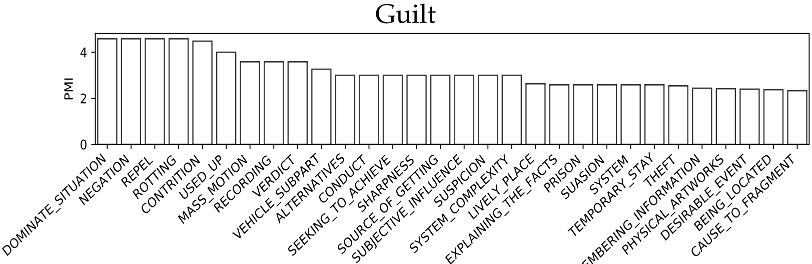
Only some of such frames overlap with the emotional output of the binary setting in Chapter 6, like: JUDGMENT_COMMUNICATION (associated to *anger*); EXECUTION, VERDICT (see *disgust* in Figure B.3); BODY_MARK and MAKING_FACES (*fear* and *joy* in the figure); ROTTING, CONTRITION (*guilt*); IMPROVEMENT_OR_DECLINE, SUBJECTIVE_INFLUENCE (*pride*); EMOTIONS_BY_STIMULUS (*relief*); IMPROVEMENT_OR_DECLINE, COMMUNICATION_NOISE (*sadness*); MANIPULATE_INTO_DOING (*shame*); JUST_FOUND_OUT, EXPERIENCER_OBJ (*surprise*); SUASION (*trust*). The other frames that are here associated to specific emotions either belong to contextually-determined frames in COCA (e.g., AGRICULTURE) or are strongly associated to the neutral label (e.g., CONTAINING). This reflects the difference in the distribution of labels between the two resources (e.g., in crowd-enVENT there is not a balanced number of neutral instances and instances corresponding to an emotion).

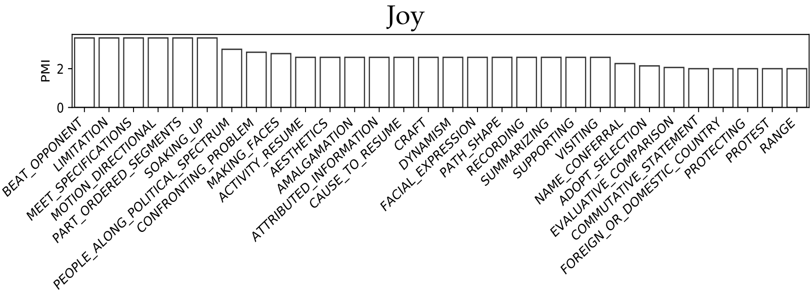

Anger
PMI
0
2
ARMOR
ARTIFICIALITY
BESIEGING
CAUSE_EMOTION
FAIRNESS_EVALUATION
HAVING_OR_LACKING_ACCESS
MOTION_NOISE
NOTIFICATION_OF_CHARGES
PREVARICATION
TRENDINESS
WAGERING
AFFIRM_OR_DENY
GRINDING
MAKE_COGNITIVE_CONNECTION
SOCIAL_INTERACTION_EVALUATION
AGREE_OR_REFUSE_TO_ACT
ADDUCING
BEING_IN_EFFECT
DOMAIN
EXPLAINING_THE_FACTS
FEIGNING
FRUGALITY
OMEN
PATTERN
QUANTITY
SUMMARIZING
JUDGMENT
DOCUMENTS
REASONING
JUDGMENT_COMMUNICATION

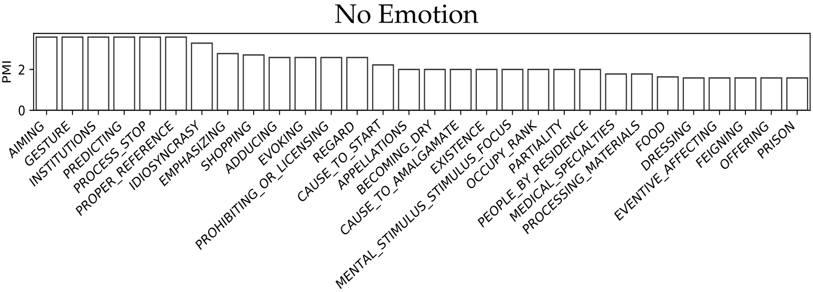

Boredom
PMI
0
2
ACCOMPANIMENT
CHANGE_EVENT_TIME
CHANGE_OF_PHASE
COMMUNICATION_MANNER
LOCALE_BY_EVENT
WAITING
AMALGAMATION
BEING_ATTACHED
BEING_OBLIGATORY
COMING_TO_BE
CONQUERING
CRAFT
HISTORY
QUANTITY
MENTAL_STIMULUS_STIMULUS_FOCUS
BEHIND_THE_SCENES
ATTENDING
BEING_RELEVANT
STATE_CONTINUE
ATTACHING
POSTURE
DISCUSSION
READING_ACTIVITY
ALTERNATIVES
BECOMING_DRY
CONNECTORS
PRACTICE
RANGE
SHARPNESS
SOURCE_OF_GETTING

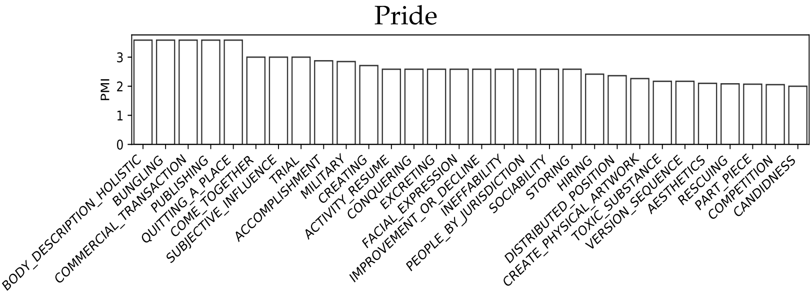

Disgust
PMI
0
2
AGRICULTURE
APPOINTING
CAUSE_FLUIDIC_MOTION
COLONIZATION
CONTAINING
EXECUTION
IMPRESSION
PROCESS_START
PROMINENCE
SEX
THWARTING
WEALTHINESS
HAVE_AS_REQUIREMENT
HIT_TARGET
REWARDS_AND_PUNISHMENTS
ACTIVITY_PAUSE
CAUSE_CHANGE_OF_PHASE
EXCRETING
FALL_ASLEEP
GUILT_OR_INNOCENCE
INGREDIENTS
MASS_MOTION
PROHIBITING_OR_LICENSING
STRICTNESS
VERDICT
INGESTION
CHANGE_OF_LEADERSHIP
FOOD_GATHERING
DEFENDING
MEDICAL_INTERVENTION

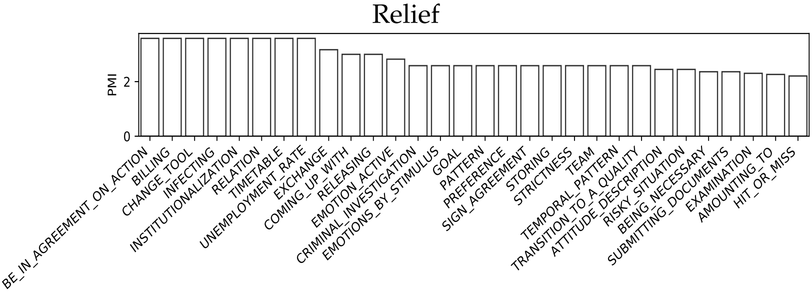

Fear
PMI
0
2
AWARENESS_STATUS
COME_DOWN_WITH
GIVING_IN
INVADING
MARGIN_OF_RESOLUTION
SMUGGLING
TRAP
UNDERGOING
FEAR
FOREIGN_OR_DOMESTIC_COUNTRY
HAIR_CONFIGURATION
RELATIONAL_NATURAL_FEATURES
SURVIVING
BODY_MARK
FRONT_FOR
CAUSE_TO_PERCEIVE
ACTIVITY_PAUSE
ARREST
CAUSE_IMPACT
CAUSE_TO_RESUME
CLOTHING_PARTS
COMING_TO_BE
DOMAIN
DYNAMISM
FALL_ASLEEP
MOVING_IN_PLACE
TRIGGERING
VISITING
LAW_ENFORCEMENT_AGENCY
SOUNDS

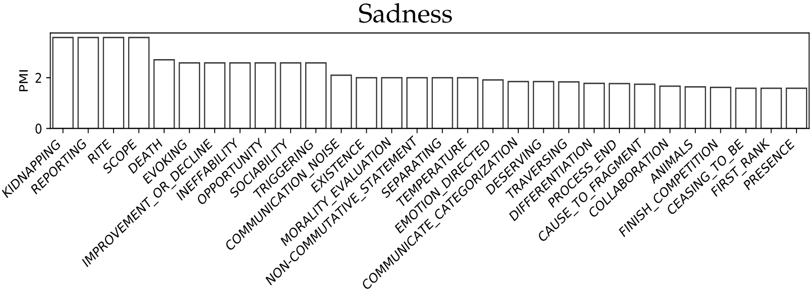

Guilt
PMI
0
2
4
DOMINATE_SITUATION
NEGATION
REPEL
ROTTING
CONTRITION
USED_UP
MASS_MOTION
RECORDING
VERDICT
VEHICLE_SUBPART
ALTERNATIVES
CONDUCT
SEEKING_TO_ACHIEVE
SHARPNESS
SOURCE_OF_GETTING
SUBJECTIVE_INFLUENCE
SUSPICION
SYSTEM_COMPLEXITY
LIVELY_PLACE
EXPLAINING_THE_FACTS
PRISON
SUASION
SYSTEM
TEMPORARY_STAY
THEFT
REMEMBERING_INFORMATION
PHYSICAL_ARTWORKS
DESIRABLE_EVENT
BEING_LOCATED
CAUSE_TO_FRAGMENT

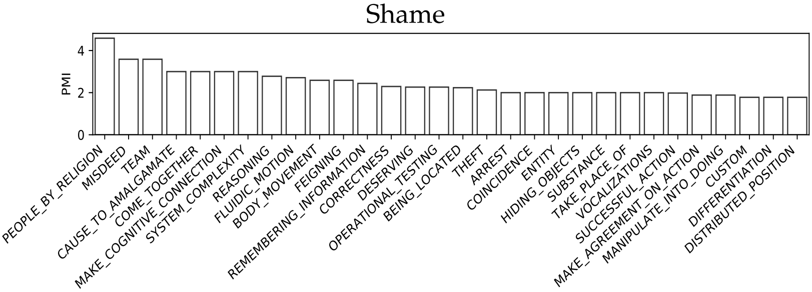

Joy
PMI
0
2
BEAT_OPPONENT
LIMITATION
MEET_SPECIFICATIONS
MOTION_DIRECTIONAL
PART_ORDERED_SEGMENTS
SOAKING_UP
PEOPLE_ALONG_POLITICAL_SPECTRUM
CONFRONTING_PROBLEM
MAKING_FACES
ACTIVITY_RESUME
AESTHETICS
AMALGAMATION
ATTRIBUTED_INFORMATION
CAUSE_TO_RESUME
CRAFT
DYNAMISM
FACIAL_EXPRESSION
PATH_SHAPE
RECORDING
SUMMARIZING
SUPPORTING
VISITING
NAME_CONFERRAL
ADOPT_SELECTION
EVALUATIVE_COMPARISON
COMMUTATIVE_STATEMENT
FOREIGN_OR_DOMESTIC_COUNTRY
PROTECTING
PROTEST
RANGE

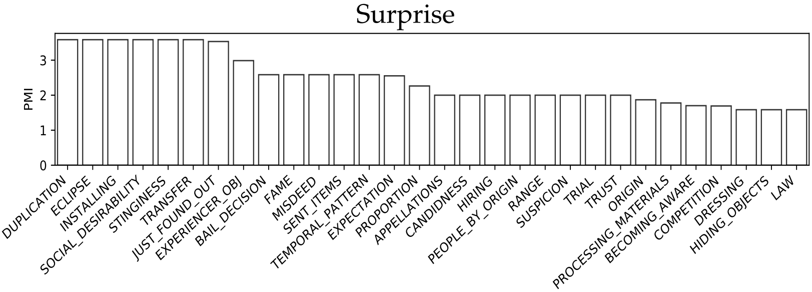

No Emotion
PMI
0
2
AIMING
GESTURE
INSTITUTIONS
PREDICTING
PROCESS_STOP
PROPER_REFERENCE
IDIOSYNCRASY
EMPHASIZING
SHOPPING
ADDUCING
EVOKING
PROHIBITING_OR_LICENSING
REGARD
CAUSE_TO_START
APPELLATIONS
BECOMING_DRY
CAUSE_TO_AMALGAMATE
EXISTENCE
MENTAL_STIMULUS_STIMULUS_FOCUS
OCCUPY_RANK
PARTIALITY
PEOPLE_BY_RESIDENCE
MEDICAL_SPECIALTIES
PROCESSING_MATERIALS
FOOD
DRESSING
EVENTIVE_AFFECTING
FEIGNING
OFFERING
PRISON


Pride
PMI
0
1
2
3
BODY_DESCRIPTION_HOLISTIC
BUNGLING
COMMERCIAL_TRANSACTION
PUBLISHING
QUITTING_A_PLACE
COME_TOGETHER
SUBJECTIVE_INFLUENCE
TRIAL
ACCOMPLISHMENT
MILITARY
CREATING
ACTIVITY_RESUME
CONQUERING
EXCRETING
FACIAL_EXPRESSION
IMPROVEMENT_OR_DECLINE
INEFFABILITY
PEOPLE_BY_JURISDICTION
SOCIABILITY
STORING
HIRING
DISTRIBUTED_POSITION
CREATE_PHYSICAL_ARTWORK
TOXIC_SUBSTANCE
VERSION_SEQUENCE
AESTHETICS
RESCUING
PART_PIECE
COMPETITION
CANDIDNESS


Relief


Sadness


Shame


Surprise

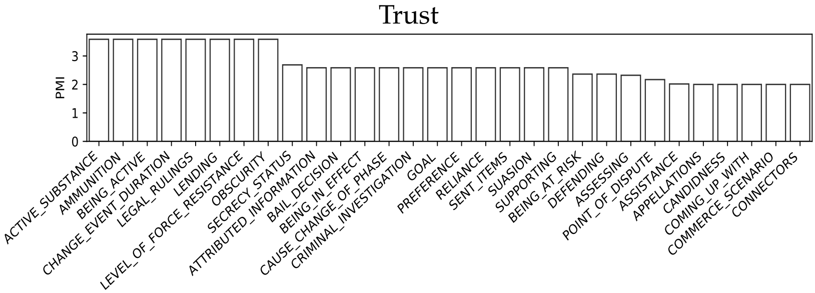


Figure B.3: The 30 frames with the highest PMI values across emotions in crowd-enVENT.

Bibliography

Abdul-Mageed, M. and Ungar, L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics. Cited on pages 45 and 219.

Akhtar, S., Ghosal, D., Ekbal, A., Bhattacharyya, P., and Kurohashi, S. (2019). All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*, pages 1–1. Cited on page 46.

Alfrjani, R., Osman, T., and Cosma, G. (2016). A new approach to ontology-based semantic modelling for opinion mining. In *2016 UKSim-AMSS 18th International Conference on Computer Modelling and Simulation (UKSim)*, pages 267–272. IEEE. Cited on page 37.

Allaway, E. and McKeown, K. (2021). A unified feature representation for lexical connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2145–2163, Online. Association for Computational Linguistics. Cited on page 58.

Allport, F. H. (1924). *Social Psychology*. Boston, Houghton Mifflin. Cited on page 20.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics. Cited on pages 35, 44, 124, 141, 166, 194, and 195.

Alvarez-Gonzalez, N., Kaltenbrunner, A., and Gómez, V. (2021). Uncovering the limits of text-based emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 3.

Alzu'bi, S., Badarneh, O., Hawashin, B., Al-Ayyoub, M., Alhindawi, N., and Jararweh, Y. (2019). Multi-label emotion classification for arabic tweets. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 499–504. IEEE. Cited on page 39.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In Matoušek, V. and Mautner, P., editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg. Cited on pages 35, 44, 139, and 166.

Arnold, M. B. (1960). *Emotion and Personality Volume 1 Psychological Aspects, and Volume 2: Neurological and Physiological Aspects*. Columbia University Press, New York, US. Cited on page 23.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24. Cited on page 229.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596. Cited on page 124.

Averill, J. R. (1980). A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier. Cited on page 19.

Bach, E. (1986). The algebra of events. *Linguistics and philosophy*, pages 5–16. Cited on page 61.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Cited on page 41.

Baker, C. F., Ellsworth, M., and Erk, K. (2007). Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth*

International Workshop on Semantic Evaluations (SemEval-2007), pages 99–104. Cited on page 197.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada. Cited on pages 63 and 189.

Balabantaray, R. C., Mohammad, M., and Sharma, N. (2012). Multi-class twitter emotion elassification: A new approach. *International Journal of Applied Information Systems*, 4(1):48–53. Cited on pages 30 and 39.

Balahur, A. and Hermida, J. M. (2012). Extending the EmotiNet knowledge base to improve the automatic detection of implicitly expressed emotions from text. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1207–1214, Istanbul, Turkey. European Language Resources Association (ELRA). Cited on page 37.

Balahur, A., Hermida, J. M., and Montoyo, A. (2012). Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1):88–101. Cited on pages 42, 43, and 75.

Balahur, A., Hermida, J. M., Montoyo, A., and Muñoz, R. (2011). EmotiNet: A knowledge base for emotion detection in text built on the appraisal theories. In Muñoz, R., Montoyo, A., and Métais, E., editors, *Natural Language Processing and Information Systems*, pages 27–39, Berlin, Heidelberg. Springer Berlin Heidelberg. Cited on pages 71 and 75.

Balahur, A. and Tanev, H. (2016). Detecting implicit expressions of affect from text using semantic knowledge on common concept properties. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA). Cited on page 219.

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages

52–60, Jeju, Korea. Association for Computational Linguistics. Cited on page 159.

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., and Curran, J. R. (2009). Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics. Cited on page 5.

Banea, C., Mihalcea, R., and Wiebe, J. (2008). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Cited on page 159.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Cited on page 50.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics. Cited on page 158.

Barnes, J. and Klinger, R. (2019). Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. *Journal of Artificial Intelligence Research*, 66:691–742. Cited on page 159.

Barnes, J., Lambert, P., and Badia, T. (2016). Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan. The COLING 2016 Organizing Committee. Cited on page 114.

Barnes, J., Oberlaender, L. A. M., Troiano, E., Kutuzov, A., Buchmann, J., Agerri, R., Øvrelid, L., and Velldal, E. (2022). SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics. Cited on page 31.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics. Cited on page 158.

Barrett, L. F. (1998). Discrete emotions or dimensions? the role of valence focus and arousal focus. *Cognition & Emotion*, 12(4):579–599. Cited on page 25.

Barrett, L. F. (2005). Feeling is perceiving: Core affect and conceptualization in the experience of emotion. *Emotion and consciousness*, pages 255–284. Cited on page 20.

Barrett, L. F. (2017). *How Emotions Are Made*. Houghton Mifflin Harcourt, New York, USA. Cited on pages 19 and 20.

Barrett, L. F. and Russell, J. A. (2015). *The psychological construction of emotion*. Guilford Publications, New York. Cited on page 19.

Barzilay, R. and Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 16–23. Cited on pages 158 and 162.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics. Cited on page 158.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. (2021). We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics. Cited on pages 125 and 150.

Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725. Cited on page 149.

Baziotis, C., Nikolaos, A., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S., and Potamianos, A. (2018). NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana. Association for Computational Linguistics. Cited on page 163.

Beckmann, N. and Wood, R. E. (2017). Dynamic personality science. integrating between-person stability and within-person change. *Frontiers in psychology*, 8:1486. Cited on page 54.

Bedford, E. (1957). Emotions. *Proceedings of the Aristotelian Society*, 57:281–304. Cited on page 26.

Bègue, I., Vaessen, M., Hofmeister, J., Pereira, M., Schwartz, S., and Vuilleumier, P. (2018). Confidence of emotion expression recognition recruits brain regions outside the face perception network. *Social Cognitive and Affective Neuroscience*, 14(1):81–95. Cited on pages 28 and 139.

Beigman Klebanov, B., Beigman, E., and Diermeier, D. (2008). Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee. Cited on page 125.

Beinborn, L., Zesch, T., and Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530. Cited on page 150.

Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2):145–204. Cited on pages 53 and 180.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166. Cited on page 40.

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71. Cited on page 39.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047, Austin, Texas. Association for Computational Linguistics. Cited on page 29.

Bestgen, Y. and Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior research methods*, 44(4):998–1006. Cited on page 45.

Bhardwaj, V., Passonneau, R., Salleb-Aouissi, A., and Ide, N. (2010). Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55, Uppsala, Sweden. Association for Computational Linguistics. Cited on page 149.

Bhowmick, P. K., Basu, A., and Mitra, P. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65, Manchester, UK. Coling 2008 Organizing Committee. Cited on page 35.

Bhowmick, P. K., Basu, A., and Mitra, P. (2010). Determining reliability of subjective and multi-label emotion annotation through novel fuzzy agreement measure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). Cited on page 152.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press. Cited on page 55.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer. Cited on page 41.

Bobicev, V. and Sokolova, M. (2017). Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria. INCOMA Ltd. Cited on pages 6 and 35.

Boldrini, E., Balahur, A., Martínez-Barco, P., and Montoyo, A. (2010). Emotiblog: A finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the fourth linguistic annotation workshop*, pages 1–10. Cited on page 47.

Bostan, L. A. M., Kim, E., and Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association. Cited on pages 71, 140, and 194.

Bostan, L. A. M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on pages 31, 58, 115, 141, 166, 192, and 193.

Bouhuys, A. L., Bloem, G. M., and Groothuis, T. G. (1995). Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects. *Journal of affective disorders*, 33(4):215–226. Cited on page 28.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59. Cited on page 20.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida, Gainesville, Florida, USA. Cited on page 45.

Bradley, M. M. and Lang, P. J. (2007). Affective norms for english text (anet): Affective ratings of text and instruction manual. *Techical Report. D-1, University of Florida, Gainesville, FL.* Cited on page 46.

Brennan, M., Afroz, S., and Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22. Cited on pages 53 and 54.

Briakou, E., Agrawal, S., Tetreault, J., and Carpuat, M. (2021a). Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on pages 50 and 178.

Briakou, E., Agrawal, S., Zhang, K., Tetreault, J., and Carpuat, M. (2021b). A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 58–67, Online. Association for Computational Linguistics. Cited on pages 50 and 178.

Briesemeister, B. B., Kuchinke, L., and Jacobs, A. M. (2011). Discrete emotion norms for nouns: Berlin affective word list (DENN–BAWL). *Behavior Research Methods*, 43(2):441. Cited on page 72.

Briot, J.-P., Hadjeres, G., and Pachet, F. (2020). *Deep learning techniques for music generation*. Springer. Cited on page 48.

Bucholtz, M. (2006). Word up: Social meanings of slang in california youth culture. *A cultural approach to interpersonal communication: Essential readings*, 243:267. Cited on page 54.

Buckley, C. (1993). The importance of proper weighting methods. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. Cited on page 198.

Buechel, S. and Hahn, U. (2016). Emotion analysis as a regression problem – dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122, The Hague, The Netherlands. Cited on page 45.

Buechel, S. and Hahn, U. (2017a). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics. Cited on pages 3, 42, and 46.

Buechel, S. and Hahn, U. (2017b). Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics. Cited on pages 34, 46, 70, 125, and 128.

Buechel, S., Hellrich, J., and Hahn, U. (2016). Feelings from the Past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 54–61, Osaka, Japan. The COLING 2016 Organizing Committee. Cited on page 45.

Buechel, S., Modersohn, L., and Hahn, U. (2021). Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 46.

Buechel, S., Rücker, S., and Hahn, U. (2020). Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics. Cited on pages 46, 71, and 159.

Buechner, V. L., Maier, M. A., Lichtenfeld, S., and Elliot, A. J. (2015). Emotion expression and color: Their joint influence on perceived attractiveness and social position. *Current Psychology*, 34(2):422–433. Cited on page 1.

Cai, J., He, S., Li, Z., and Zhao, H. (2018). A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2753–2765, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on pages 64 and 65.

Cai, R. and Lapata, M. (2019). Syntax-aware semantic role labeling without parsing. *Transactions of the Association for Computational Linguistics*, 7:343–356. Cited on page 65.

Calvo, R. A. and Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543. Cited on pages 38, 45, and 46.

Canales, L. and Martínez-Barco, P. (2014). Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador. Association for Computational Linguistics. Cited on page 38.

Cannon, W. B. (1927). The James-Lange theory of emotions: a critical examination and an alternative theory. *The American Journal of Psychology*, 39:106–124. Place: US Publisher: Univ of Illinois Press. Cited on page 17.

Canty, A. and Ripley, B. D. (2021). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28. Cited on pages 104 and 133.

Carnap, R. (1988). *Meaning and necessity: a study in semantics and modal logic*, volume 30. University of Chicago Press. Cited on page 60.

Casel, F., Heindl, A., and Klinger, R. (2021). Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers. Cited on page 72.

Cevher, D., Zepf, S., and Klinger, R. (2019). Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 79–90, Erlangen, Germany. German Society for Computational Linguistics & Language Technology. Cited on page 44.

Chakrabarty, T., Hidey, C., and Muresan, S. (2021). ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4958–4971, Online. Association for Computational Linguistics. Cited on pages 57, 58, and 59.

Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., and Demartini, G. (2017). Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*. Cited on page 229.

Chen, E., Shi, L., and Hu, D. (2008). Probabilistic model for syntactic and semantic dependency parsing. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 263–267, Manchester, England. Coling 2008 Organizing Committee. Cited on page 64.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics. Cited on pages 31 and 159.

Cheng, P., Min, M. R., Shen, D., Malon, C., Zhang, Y., Li, Y., and Carin, L. (2020). Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online. Association for Computational Linguistics. Cited on pages 50 and 51.

Chepenik, L. G., Cornew, L. A., and Farah, M. J. (2007). The influence of sad mood on cognition. *Emotion*, 7(4):802. Cited on page 28.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29. Cited on page 198.

Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558. Cited on pages 34 and 142.

Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F. (2017). A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics. Cited on page 31.

Clark, E. A., Kessinger, J., Duncan, S. E., Bell, M. A., Lahne, J., Gallagher, D. L., and O'Keefe, S. F. (2020). The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review. *Frontiers in Psychology*, 11. Cited on page 18.

Clore, G. L. and Ortony, A. (2013). Psychological Construction in the OCC Model of Emotion. *Emotion Review*, 5(4):335–343. Cited on pages 22, 46, and 83.

Clore, G. L., Ortony, A., and Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53(4):751–766. Cited on pages 43 and 188.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46. Cited on pages 34, 71, and 142.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537. Cited on pages 39 and 65.

Coseriu, E. and Geckeler, H. (1974). Linguistics and semantics. *Current trends in linguistics*, 12(pt 1):103–171. Cited on page 60.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press. Cited on pages 60 and 214.

Dai, N., Liang, J., Qiu, X., and Huang, X. (2019). Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics. Cited on page 50.

Danisman, T. and Alpkocak, A. (2008). Feeler: Emotion classification of text using vector space model. In *AISB 2008 convention communication, interaction and social intelligence*, volume 1, pages 53–59. Cited on page 47.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray. Cited on page 27.

Das, D. and Bandyopadhyay, S. (2011). Analyzing emotional statements – roles of general and physiological variables. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 59–67, Chiang Mai, Thailand. Asian Federation of Natural Language Processing. Cited on page 47.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, California. Association for Computational Linguistics. Cited on page 65.

Das, D. and Smith, N. A. (2011). Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting* *of the Association for Computational Linguistics: Human Language Technologies*, pages 1435–1444, Portland, Oregon, USA. Association for Computational Linguistics. Cited on page 65.

Davidson, D. (1967). The logical form of action sentences. *The logic of decision and action*, pages 81–95. Cited on page 61.

Davies, M. (2015). Corpus of Contemporary American English (COCA). Cited on pages 128 and 192.

Davis, A. R. (2019). Thematic roles. In *Semantics: Lexical Structures and Adjectives*, pages 99–125. Mounton de Gruyter. Cited on page 61.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge. ISBN 0-521-57391-2. Cited on pages 104 and 133.

Dayanik, E., Vu, T., and Padó, S. (2022). Bias identification and attribution in NLP models with regression and effect sizes. *Northern European Journal of Language Technology*, 8(1). Cited on page 150.

De Bruyne, L., Clercq, O. D., and Hoste, V. (2021). Prospects for dutch emotion detection: Insights from the new emotionl dataset. *Computational Linguistics in the Netherlands Journal*, 11:231–. Cited on page 72.

De Mattei, L., Cafagna, M., Dell'Orletta, F., and Nissim, M. (2020). Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6709–6717, Marseille, France. European Language Resources Association. Cited on page 177.

De Saussure, F. (1959). *Course in General Linguistics*. Philosophical Library, New York. Cited on page 53.

De Sousa, R. (1990). *The rationality of emotion*. Mit Press. Cited on page 121.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407. Cited on page 38.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics. Cited on pages 45, 195, and 219.

Derks, D., Bos, A. E., and Von Grumbkow, J. (2008). Emoticons and online message interpretation. *Social Science Computer Review*, 26(3):379–388. Cited on page 71.

Descartes, R. (1989). *Passions of the Soul*. Hackett Publishing. First published in French in 1649. Cited on page 3.

Devillers, L., Cowie, R., Martin, J. C., Douglas-Cowie, E., Abrilian, S., and McRorie, M. (2006). Real life emotions in French and English TV video clips: An integrated annotation protocol combining continuous and discrete approaches. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA). Cited on page 29.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 40, 141, and 195.

Dewey, J. (1894). The theory of emotion. (i.) emotional attitudes. *Psychological review*, 1:553–569. Cited on page 17.

Dixon, T. (2012). "Emotion": The history of a keyword in crisis. *Emotion Review*, 4(4):338–344. Cited on page 3.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS ONE*, 6(12):e26752. Cited on page 36.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619. Cited on page 62.

Dryjański, T., Bujnowski, P., Choi, H., Podlaska, K., Michalski, K., Beksa, K., and Kubik, P. (2018). Affective natural language generation by phrase insertion. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4876–4882. Cited on pages 52 and 58.

Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462. Cited on page 18.

Dunn, J. R. and Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology*, 88(5):736. Cited on page 86.

Edmonds, D. and Sedoc, J. (2021). Multi-emotion classification for song lyrics. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 221–235, Online. Association for Computational Linguistics. Cited on page 5.

Eisenberg, N. and Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological bulletin*, 101(1):91. Cited on page 188.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In Cole, J., editor, *Nebraska Symposium on Motivation 1971*, volume 19. Lincoln University of Nebraska Press. Cited on pages 28 and 126.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200. Cited on pages 17, 19, 31, 44, and 166.

Ekman, P. (1999). *Basic Emotions*, chapter 3, pages 45–60. John Wiley & Sons, Ltd. Cited on page 17.

Ekman, P. and Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review*, 3(4):364–370. Cited on pages 17 and 31.

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto. Cited on page 27.

Ekman, P., Friesen, W. V., and Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, 39(6):1125–1134. Cited on page 27.

Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203. Cited on page 28.

Elfenbein, H. A. and Ambady, N. (2003a). Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition. *Journal of Cross-Cultural Psychology*, 34(1):92–110. Cited on page 28.

Elfenbein, H. A. and Ambady, N. (2003b). When familiarity breeds accuracy: cultural exposure and facial emotion recognition. *Journal of personality and social psychology*, 85(2):276. Cited on page 28.

Ellsworth, P. C. and Smith, C. A. (1988). From appraisal to emotion: Differences among unpleasant feelings. *Motivation and emotion*, 12(3):271–302. Cited on page 23.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211. Cited on page 39.

Erk, K. and Padó, S. (2006). Shalmaneser - a toolchain for shallow semantic parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA). Cited on page 65.

Esuli, A., Sebastiani, F., and Urciuoli, I. (2008). Annotating expressions of opinion and emotion in the Italian content annotation bank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Cited on page 5.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74. Cited on page 193.

Evens, M. W., Litowitz, B. C., Markowitz, J. A., Smith, R. N., and Werner, O. (1980). Lexical-semantic relations: A comparative survey. *Linguistic Research*, pages 187–219. Cited on page 60.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics. Cited on page 163.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874. Cited on page 115.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics. Cited on pages 191 and 217.

Fei, H., Zhang, M., Li, B., and Ji, D. (2021). End-to-end semantic role labeling with neural transition-based model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12803–12811. Cited on page 65.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics. Cited on pages 2, 33, 42, 44, 71, 189, and 219.

Feldman Barrett, L. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46. Cited on page 20.

Feldman Barrett, L. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(11):1833. Cited on page 20.

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*. Language, speech, and communication. MIT Press, Cambridge, Mass. Cited on page 43.

Felt, C. and Riloff, E. (2020). Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145, Online. Association for Computational Linguistics. Cited on page 181.

Fernández-González, D. and Gómez-Rodríguez, C. (2020). Transition-based semantic dependency parsing with pointer networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 7035–7046, Online. Association for Computational Linguistics. Cited on page 65.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea. Cited on pages 62 and 189.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, 6(2):222–254. Cited on pages 9 and 189.

Fillmore, C. J. and Baker, C. F. (2000). Framenet: Frame semantics meets the corpus. Poster session at the Annual Meeting of the Linguistic Society of America. Cited on page 63.

Fillmore, C. J. et al. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, 280(1):20–32. Cited on page 219.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382. Cited on pages 34, 142, and 144.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics. Cited on page 124.

Fossum, V. and Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics. Cited on page 41.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50. Cited on page 177.

Friedman, B., Kahn Jr., P. H., and Borning, A. (2006). Value sensitive design and information systems. *Human-computer interaction in management information systems: Foundations*, pages 348–372. Cited on page 229.

Frijda, N. H. (1996). Passions: Emotion and socially consequential behavior. *Emotion: Interdisciplinary Perspectives*, 1:1–17. Cited on page 17.

Frijda, N. H., Ortony, A., Sonnemans, J., and Clore, G. L. (1992). The complexity of intensity: Issues concerning the structure of emotion intensity. *Emotion*, pages 60–89. Cited on page 19.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). Cited on pages 51 and 52.

Fussell, S. R. (2002). The verbal communication of emotion: Introduction and overview. In *The verbal communication of emotions*, pages 9–24. Psychology Press. Cited on page 1.

Galeshchuk, S., Qiu, J., and Jourdan, J. (2019). Sentiment analysis for multilingual corpora. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 120–125, Florence, Italy. Association for Computational Linguistics. Cited on page 31.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170. Cited on page 48.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423. Cited on page 48.

Gendron, M. and Feldman Barrett, L. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1(4):316–339. Cited on page 16.

Ghazi, D., Inkpen, D., and Szpakowicz, S. (2015). Detecting emotion stimuli in emotion-bearing sentences. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 152–165, Cham. Springer International Publishing. Cited on pages 66, 141, 192, and 195.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., and Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics. Cited on pages 2 and 183.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288. Cited on page 61.

Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. (2008). Identifying emotional characteristics from short blog texts. In *30th Annual Conference of the Cognitive Science Society*, pages 2237–2242. Cited on page 38.

Gimpel, K., Batra, D., Dyer, C., and Shakhnarovich, G. (2013). A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Seattle, Washington, USA. Association for Computational Linguistics. Cited on page 163.

Goecke, D., Stührenberg, M., and Witt, A. (2008). Influence of text type and text length on anaphoric annotation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Cited on page 5.

Golightly, C. L. (1953). The jameslange theory: A logical postmortem. *Philosophy of Science*, 20(4):286–299. Cited on page 19.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics. Cited on page 41.

Gosling, S. D., Rentfrow, P. J., and Swann Jr., W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528. Cited on page 87.

Goswamy, T., Singh, I., Barkati, A., and Modi, A. (2020). Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics. Cited on pages 3 and 183.

Grice, P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax And Semantics*. Academic Press, New York. Cited on pages 5 and 188.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928. Cited on page 37.

Guibon, G., Labeau, M., Flamein, H., Lefeuvre, L., and Clavel, C. (2021). Few-shot emotion recognition in conversation with sequential prototypical networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6858–6870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 3.

Haider, T., Eger, S., Kim, E., Klinger, R., and Menninghaus, W. (2020). PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association. Cited on pages 35, 111, 124, and 140.

Hakak, N. M., Mohd, M., Kirmani, M., and Mohd, M. (2017). Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, pages 397–402. Cited on page 36.

Hall, J. A., Mast, M. S., and West, T. V. (2016). Accurate interpersonal perception: Many traditions, one topic. In *The Social Psychology of Perceiving Others Accurately*, page 3–22. Cambridge University Press. Cited on page 28.

Hall, J. A. and Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, 4(2):201–206. Cited on page 126.

Hampson, E., van Anders, S. M., and Mullin, L. I. (2006). A female advantage in the recognition of emotional facial expressions: Test of an evolutionary hypothesis. *Evolution and Human Behavior*, 27(6):401–416. Cited on page 126.

Hartmann, S., Kuznetsov, I., Martin, T., and Gurevych, I. (2017). Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th*

Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 471–482, Valencia, Spain. Association for Computational Linguistics. Cited on page 193.

He, H., Grissom II, A., Morgan, J., Boyd-Graber, J., and Daumé III, H. (2015). Syntax-based rewriting for simultaneous machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 55–64, Lisbon, Portugal. Association for Computational Linguistics. Cited on page 158.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics. Cited on page 65.

Helbig, D., Troiano, E., and Klinger, R. (2020). Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 41–50, Online. Association for Computational Linguistics. Cited on pages 56, 58, 59, 157, 161, 162, 177, and 178.

Higginbotham, J. (1985). On semantics. *Linguistic inquiry*, 16(4):547–593. Cited on page 61.

Higginbotham, J. (2000). On events in linguistic semantics. *Speaking of events*, 49(80):49–79. Cited on page 61.

Hobbs, J. R. and Gordon, A. S. (2011). The deep lexical semantics of emotions. In *Affective Computing and Sentiment Analysis*, pages 27–34. Springer. Cited on page 43.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. Cited on page 40.

Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., and Traue, H. C. (2010). Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men. *Acta Psychologica*, 135(3):278–283. Cited on page 28.

Hofmann, J., Troiano, E., and Klinger, R. (2021). Emotion-aware, emotion-agnostic, or automatic: Corpus creation strategies to obtain cognitive event appraisal annotations. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 160–170, Online. Association for Computational Linguistics.

Hofmann, J., Troiano, E., Sassenberg, K., and Klinger, R. (2020). Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. Cited on page 38.

Hogan, R., DeSoto, C. B., and Solano, C. (1977). Traits, tests, and personality research. *American Psychologist*, 32(4):255. Cited on page 153.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics. Cited on page 40.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. Cited on page 198.

Hu, M. and Liu, B. (2006). Opinion extraction and summarization on the web. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, page 1621–1624. AAAI Press. Cited on page 56.

Hu, Z., Lee, R. K.-W., and Aggarwal, C. C. (2020). Text style transfer: A review and experiment evaluation. *arXiv preprint arXiv:2010.12742*. Cited on page 51.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2017). Toward controlled generation of text. *Proceedings of Machine Learning Research*, 70:1587–1596. Cited on page 52.

Huang, C., Zaïane, O., Trabelsi, A., and Dziri, N. (2018). Automatic dialogue generation with expressed emotions. In *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 49–54, New Orleans, Louisiana. Association for Computational Linguistics. Cited on pages 3 and 52.

Hubscher-Davidson, S. (2017). *Translation and Emotion: A psychological perspective*. Routledge. Cited on page 159.

Ide, N., Baker, C. F., Fellbaum, C., and Passonneau, R. J. (2010). The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 conference short papers*, pages 68–73. Cited on page 197.

Ikeda, S. (2020). Influence of color on emotion recognition is not bidirectional: An investigation of the association between color and emotion using a stroop-like task. *Psychological reports*, 123(4):1226–1239. Cited on page 27.

Imbir, K. K. (2017). The affective norms for polish short texts (anpst) database properties and impact of participants' population and sex on affective ratings. *Frontiers in Psychology*, 8:855. Cited on page 46.

Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts, New York. Cited on page 17.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., and Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current biology*, 19(18):1543–1548. Cited on page 126.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., and Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522. Cited on page 159.

Jackson, P. (1986). *Introduction to expert systems*. Addison-Wesley Pub. Co., Reading, MA. Cited on page 36.

Jacobs, G. and Hoste, V. (2021). Fine-grained implicit sentiment in financial news: Uncovering hidden bulls and bears. *Electronics*, 10(20):2554. Cited on page 75.

Jacobs, G. and Hoste, V. (2022). Sentivent: enabling supervised information extraction of company-specific events in economic and financial news. *Language Resources and Evaluation*, 56(1):225–257. Cited on page 75.

Jacobson, P. I. (2014). *Compositional semantics: An introduction to the syntax/semantics interface*. Oxford University Press. Cited on page 60.

James, W. (1890). *The Principles of Psychology*. New York, Holt. Cited on page 26.

James, W. (1894). Discussion: The physical basis of emotion. *Psychological review*, 1(5):516–529. Cited on page 19.

Jamison, E. and Gurevych, I. (2015). Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics. Cited on page 124.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In *EuroVis (STARs)*, pages 83–103. Cited on page 193.

Janssen, T. M. and Partee, B. H. (1997). Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier. Cited on page 60.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. (2022). Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205. Cited on pages 6 and 51.

Jin, Z., Jin, D., Mueller, J., Matthews, N., and Santus, E. (2019). IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics. Cited on page 51.

John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. (2019). Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics. Cited on page 51.

Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India. Cited on page 62.

Juslin, P. N. and Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, 1(4):381. Cited on pages 28 and 127.

Kajava, K., Ohman, E., Hui, P., and Tiedemann, J. (2020). Emotion preservation in translation: Evaluating datasets for annotation projection. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings. Cited on page 160.

Kamps, J. and Marx, M. (2001). Words with attitude. *1st International WordNet Conference*. Cited on page 59.

Kang, D., Gangal, V., and Hovy, E. (2019). (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics. Cited on pages 181 and 184.

Kang, D. and Hovy, E. (2021). Style is NOT a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387, Online. Association for Computational Linguistics. Cited on pages 54 and 182.

Kao, E. C.-C., Liu, C.-C., Yang, T.-H., Hsieh, C.-T., and Soo, V.-W. (2009). Towards text-based emotion detection a survey and possible improvements. In *2009 International Conference on Information Management and Engineering*, pages 70–74. IEEE. Cited on page 36.

Kasai, J., Friedman, D., Frank, R., Radev, D., and Rambow, O. (2019). Syntax-aware neural semantic role labeling with supertags. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 701–709, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 65.

Kavanaugh, R. D., Zimmerberg, B., and Fein, S. (1996). *Emotion: Interdisciplinary Perspectives*. Psychology Press. Cited on page 16.

Kelly, K. J. and Metcalfe, J. (2011). Metacognition of emotional face recognition. *Emotion*, 11(4):896. Cited on page 28.

Kim, E. and Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on pages 35, 66, and 71.

Kim, E. and Klinger, R. (2019). Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 30 and 31.

Kim, E., Padó, S., and Klinger, R. (2017). Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics. Cited on page 44.

Klineberg, O. (1940). *Social Psychology*. New York, Holt. Cited on page 20.

Klinger, R., de Clercq, O., Mohammad, S. M., and Balahur, A. (2018). IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics. Cited on pages 3, 33, 44, 45, 71, and 91.

Köper, M., Kim, E., and Klinger, R. (2017). IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark. Association for Computational Linguistics. Cited on page 45.

Kratzer, A. and Heim, I. (1998). *Semantics in generative grammar*, volume 1185. Blackwell Oxford. Cited on page 60.

Krcadinac, U., Pasquier, P., Jovanovic, J., and Devedzic, V. (2013). Synesketch: An open source library for sentence-based emotion recognition. *IEEE Transactions on Affective Computing*, 4(3):312–325. Cited on page 38.

Krishna, K., Wieting, J., and Iyyer, M. (2020). Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics. Cited on pages 52 and 184.

Kuznetsov, I. and Gurevych, I. (2018). From text to lexicon: Bridging the gap between word embeddings and lexical resources. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 233–244, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on page 191.

Lai, C.-T., Hong, Y.-T., Chen, H.-Y., Lu, C.-J., and Lin, S.-D. (2019). Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3579–3584, Hong Kong, China. Association for Computational Linguistics. Cited on page 51.

Lample, G., Subramanian, S., Smith, E. M., Denoyer, L., Ranzato, M., and Boureau, Y. (2019). Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. Cited on pages 52 and 57.

Lamprinidis, S., Bianchi, F., Hardt, D., and Hovy, D. (2021). Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online. Association for Computational Linguistics. Cited on page 72.

Landis, C. (1924). Studies of emotional reactions. ii. general behavior and facial expression. *Journal of Comparative Psychology*, 4(5):447–509. Cited on page 20.

Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., and Althoff, J. (2010). Presenting the venec corpus:

Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect*, pages 53–57. European Language Resources Association Paris, France. Cited on page 76.

Lausen, A. and Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1):1–17. Cited on pages 28 and 139.

Lawrence, K., Campbell, R., and Skuse, D. (2015). Age, gender, and puberty influence the development of facial emotion recognition. *Frontiers in psychology*, 6:761. Cited on page 27.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791. Cited on page 38.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72. Cited on page 54.

Lee, J. (2020). Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 195–204, Dublin, Ireland. Association for Computational Linguistics. Cited on page 51.

Lee, J., Xie, Z., Wang, C., Drach, M., Jurafsky, D., and Ng, A. (2019). Neural text style transfer via denoising and reranking. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 74–81, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 52.

Leppänen, J. M. (2006). Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings. *Current opinion in psychiatry*, 19(1):34–39. Cited on page 27.

Lewis, M. D. (2001). Personal pathways in the development of appraisal. *Appraisal processes in emotion: Theory, methods, research*, pages 205–220. Cited on page 86.

Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics. Cited on pages 51 and 177.

Li, M., Wang, D., Lu, Q., and Long, Y. (2016a). Event based emotion classification for news articles. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 153–162, Seoul, South Korea. Cited on page 5.

Li, S., Lu, Q., Zhao, T., Liu, P., and Li, H. (2010). Combining constituent and dependency syntactic views for Chinese semantic role labeling. In *Coling 2010: Posters*, pages 665–673, Beijing, China. Coling 2010 Organizing Committee. Cited on page 65.

Li, S., Xu, J., Zhang, D., and Zhou, G. (2016b). Two-view label propagation to semi-supervised reader emotion classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2647–2655, Osaka, Japan. The COLING 2016 Organizing Committee. Cited on page 30.

Li, Y., Li, C., Zhang, Y., Li, X., Zheng, G., Carin, L., and Gao, J. (2020). Complementary auxiliary classifiers for label-conditional text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8303–8310. Cited on page 58.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing. Cited on pages 42, 44, 141, 189, and 195.

Li, Z., Zhao, H., He, S., and Cai, J. (2021). Syntax role for neural semantic role labeling. *Computational Linguistics*, 47(3):529–574. Cited on page 64.

Liew, J. S. Y. (2014). Expanding the range of automatic emotion detection in microblogging text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–44, Gothenburg, Sweden. Association for Computational Linguistics. Cited on page 5.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Cited on page 50.

Lin, K., Liu, M.-Y., Sun, M.-T., and Kautz, J. (2020). Learning to generate multiple style transfer outputs for an input sentence. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 10–23, Online. Association for Computational Linguistics. Cited on pages 51 and 52.

Ling, H. S., Bali, R., and Salam, R. A. (2006). Emotion detection using keywords spotting and semantic network ieee icoci 2006. In *2006 International Conference on Computing & Informatics*, pages 1–5. IEEE. Cited on page 38.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan&Claypool. Cited on page 75.

Liu, H., Lieberman, H., and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. Cited on page 38.

Liu, V., Banea, C., and Mihalcea, R. (2007). Grounded emotions. In *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, pages 477–483. IEEE Computer Society. Cited on pages 141 and 194.

Lohar, P., Afli, H., and Way, A. (2017). Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108(1):73–84. Cited on page 160.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press. Cited on page 32.

Luo, F., Li, P., Yang, P., Zhou, J., Tan, Y., Chang, B., Sui, Z., and Sun, X. (2019). Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022, Florence, Italy. Association for Computational Linguistics. Cited on pages 52 and 164.

Lykousas, N., Patsakis, C., Kaltenbrunner, A., and Gómez, V. (2019). Sharing emotions at scale: The vent dataset. In *Proceedings of the*

International AAAI Conference on Web and Social Media, volume 13, pages 611–619. Cited on page 33.

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., and Prabhumoye, S. (2020). Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics. Cited on pages 51 and 177.

Maienborn, C. (2011). Event semantics. *Edited by Claudia Maienborn Klaus von Heusinger*, pages 802–829. Cited on page 61.

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics. Cited on page 2.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics. Cited on pages 158 and 164.

Malmi, E., Severyn, A., and Rothe, S. (2020). Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics. Cited on page 51.

Mancini, G., Biolcati, R., Agnoli, S., Andrei, F., and Trombini, E. (2018). Recognition of facial emotional expressions among italian pre-adolescents, and their affective reactions. *Frontiers in Psychology*, 9. Cited on page 27.

Manstead, A. and Tetlock, P. E. (1989). Cognitive appraisals and emotional experience: Further evidence. *Cognition and Emotion*, 3(3):225–240. Cited on page 86.

Marcheggiani, D., Frolov, A., and Titov, I. (2017). A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 411–420, Vancouver, Canada. Association for Computational Linguistics. Cited on page 65.

Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. (2008). Special issue introduction: Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159. Cited on pages 61 and 65.

Martin, J. R. and White, P. R. (2003). *The language of evaluation*, volume 2. Springer. Cited on page 32.

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., Yrizarry, N., Loewinger, S., Uchida, H., Yee, A., et al. (2000). A new test to measure emotion recognition ability: Matsumoto and ekman's japanese and caucasian brief affect recognition test (jacbart). *Journal of Nonverbal behavior*, 24(3):179–209. Cited on page 126.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292. Cited on page 183.

Mehrabian, A. and Epstein, N. (1972). A measure of emotional empathy. *Journal of personality*. Cited on page 188.

Melzi, S., Abdaoui, A., Azé, J., Bringay, S., Poncelet, P., and Galtier, F. (2014). Patient's rationale: Patient knowledge retrieval from health forums. In *eTELEMED: eHealth, Telemedicine, and Social Medicine*. Cited on page 124.

Merity, S., Keskar, N. S., and Socher, R. (2018). Regularizing and optimizing LSTM language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. Cited on page 40.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic. Association for Computational Linguistics. Cited on page 159.

Mihalcea, R. and Strapparava, C. (2012). Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea. Association for Computational Linguistics. Cited on page 44.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Cited on page 40.

Mill, A., Allik, J., Realo, A., and Valk, R. (2009). Age-related differences in emotion recognition ability: a cross-sectional study. *Emotion*, 9(5):619–630. Cited on page 126.

Minsky, M. (2006). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster. Cited on pages 1, 229, and 230.

Mir, R., Felbo, B., Obradovich, N., and Rahwan, I. (2019). Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 50.

Mirkin, S., Nowson, S., Brun, C., and Perez, J. (2015). Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal. Association for Computational Linguistics. Cited on page 159.

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics. Cited on page 5.

Mohammad, S. (2012). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for

Computational Linguistics. Cited on pages 3, 33, 35, 44, 58, 71, 141, 166, 194, and 195.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics. Cited on pages 34, 45, 128, and 138.

Mohammad, S. and Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics. Cited on pages 44, 141, and 195.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics. Cited on pages 44 and 46.

Mohammad, S., Shutova, E., and Turney, P. (2016a). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics. Cited on page 181.

Mohammad, S., Zhu, X., and Martin, J. (2014). Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics. Cited on pages 66 and 71.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Meiselman, H., editor, *Emotion Measurement*. Elsevier. Cited on page 20.

Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278. Cited on page 122.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016b). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130. Cited on pages 159 and 177.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465. Cited on pages 42, 43, 70, and 188.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press. Cited on page 39.

Montague, R. (1970). Universal grammar. *Theoria*, 36(3):373–398. Cited on page 60.

Moors, A., Ellsworth, P. C., Scherer, K. R., and Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124. Cited on page 21.

Mortillaro, M., Meuleman, B., and Scherer, K. R. (2012). Advocating a componential appraisal model to guide emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 3(1):18–32. Cited on pages xi, 23, 26, 72, 76, and 77.

Mukherjee, J. (2005). Stylistics. *Encyclopedia of Linguistics*, pages 1043–1045. Cited on page 180.

Mukherjee, R., Naik, A., Poddar, S., Dasgupta, S., and Ganguly, N. (2021). Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2303–2307. Association for Computing Machinery, New York, NY, USA. Cited on page 46.

Munot, R. and Nenkova, A. (2019). Emotion impacts speech recognition performance. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 16–21, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 29.

Musat, C. and Trausan-Matu, S. (2010). The impact of valence shifters on mining implicit economic opinions. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 131–140. Springer. Cited on page 75.

Myers, G. E. (1969). William james's theory of emotion. *Transactions of the Charles S. Peirce Society*, 5(2):67–89. Cited on page 19.

Nangi, S. R., Chhaya, N., Khosla, S., Kaushik, N., and Nyati, H. (2021). Counterfactuals to control latent disentangled text representations for style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 40–48, Online. Association for Computational Linguistics. Cited on page 59.

Navas Alejo, I., Badia, T., and Barnes, J. (2020). Cross-lingual emotion intensity prediction. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 140–152, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on page 71.

Nelson, N. L. and Russell, J. A. (2013). Universality revisited. *Emotion Review*, 5(1):8–15. Cited on page 126.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2007). Textual affect sensing for sociable and expressive online communication. In *International Conference on Affective Computing and Intelligent Interaction*, pages 218–229. Springer. Cited on page 43.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, pages 278–281. Cited on pages 36 and 43.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics. Cited on page 166.

Ngo, A., Candri, A., Ferdinan, T., Kocon, J., and Korczynski, W. (2022). Studemo: A non-aggregated review dataset for personalized emotion recognition. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 46–55, Marseille, France. European Language Resources Association. Cited on page 229.

Nida, E. (1979). *Componential Analysis of Meaning: An Introduction to Semantic Structures*. Approaches to semiotics. Mouton De Gruyter. Cited on page 60.

Niedenthal, P. M., Halberstadt, J. B., Margolin, J., and Innes-Ker, Å. H. (2000). Emotional state and the detection of change in facial expression of emotion. *European journal of social psychology*, 30(2):211–222. Cited on pages 27 and 127.

Nussbaum, M. (2004). Emotions as judgments of value and importance. *Thinking about feeling: Contemporary philosophers on emotions*, pages 183–199. Cited on page 1.

Oatley, K. (1993). Social construction in emotions. In Lewis, M. and Haviland, J. M., editors, *Handbook of emotions*, pages 341–352. Guilford Press, New York. Cited on page 19.

Oberländer, L. A. M. and Klinger, R. (2020). Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on pages 66 and 219.

Oberländer, L. A. M., Reich, K., and Klinger, R. (2020). Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, Barcelona, Spain. Association for Computational Linguistics. Cited on pages 66, 189, and 219.

Omdahl, B. L. (1995). *Cognitive Appraisal, Emotion, and Empathy*. Mahwah, NJ: Lawrence Erlbaum. Cited on pages 72 and 188.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge university press. Cited on pages 21, 99, and 219.

Osgood, C. E., May, W. H., Miron, M. S., and Miron, M. S. (1975). *Cross-cultural universals of affective meaning*, volume 1. University of Illinois Press. Cited on page 60.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois press. Cited on page 60.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 163.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106. Cited on page 62.

Pang, R. Y. (2019). The daunting task of real-world textual style transfer auto-evaluation. *arXiv preprint arXiv:1910.03747*. Cited on page 50.

Pang, R. Y. and Gimpel, K. (2019). Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 138–147, Hong Kong. Association for Computational Linguistics. Cited on page 51.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press. Cited on page 3.

Papay, S., Klinger, R., and Padó, S. (2020). Dissecting span identification tasks with performance prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4881–4895, Online. Association for Computational Linguistics. Cited on page 150.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Cited on page 50.

Park, S., Kim, J., Ye, S., Jeon, J., Park, H. Y., and Oh, A. (2021). Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 46.

Parkinson, B., Fischer, A. H., and Manstead, A. S. (2005). *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology press. Cited on page 27.

Partee, B. (1973). Some transformational extensions of montague grammar. *Journal of Philosophical Logic*, 2(4):509–534. Cited on page 60.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR. Cited on page 40.

Pekar, V., Krkoska, M., and Staab, S. (2004). Feature weighting for co-occurrence-based classification of words. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 799–805, Geneva, Switzerland. COLING. Cited on page 198.

Peldszus, A. and Stede, M. (2013). Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria. Association for Computational Linguistics. Cited on page 149.

Pennebaker, J. W. and Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291–301. Cited on pages 54 and 181.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. Cited on pages 40 and 166.

Petrova, O. and Rodionova, M. (2016). Rendering emotional coloring in literary translation. *Procedia-Social and Behavioral Sciences*, 231:195–202. Cited on pages 114 and 159.

Pianesi, F. and Varzi, A. C. (2000). Events and event talk. *Speaking of events*, pages 3–47. Cited on page 61.

Picard, R. W. (2000). *Affective Computing*. MIT press. Cited on pages 2, 29, and 229.

Plank, B., Hovy, D., and Søgaard, A. (2014a). Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational*

Linguistics, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics. Cited on page 124.

Plank, B., Hovy, D., and Søgaard, A. (2014b). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics. Cited on pages 149 and 229.

Plutchik, R. (1970). Emotions, evolution and adaptive processes. In Arnold, M. B., editor, *Feelings and emotions*, pages 3–24. New York: Academic Press. Cited on pages 18 and 19.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4. Cited on page 18.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350. Cited on pages 17, 18, 19, and 31.

Pool, C. and Nissim, M. (2016). Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan. The COLING 2016 Organizing Committee. Cited on page 47.

Poquérusse, J., Pastore, L., Dellantonio, S., and Esposito, G. (2018). Alexithymia and autism spectrum disorder: a complex relationship. *Frontiers in psychology*, 9:1196. Cited on page 57.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics. Cited on page 44.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734. Cited on page 20.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics. Cited on pages 52, 160, and 161.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J., and Jurafsky, D. (2005). Semantic role labeling using different syntactic views. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 581–588, Ann Arbor, Michigan. Association for Computational Linguistics. Cited on page 65.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (2016). Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics. Cited on pages 3 and 46.

Prinz, J. (2005). Are emotions feelings? *Journal of consciousness studies*, 12(8-9):9–25. Cited on page 120.

Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. oxford university Press. Cited on page 1.

Punyakanok, V., Roth, D., and Yih, W.-t. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287. Cited on page 64.

Qadir, A. and Riloff, E. (2014). Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1203–1209. Association for Computational Linguistics. Cited on page 71.

Qing, C., Endriss, U., Fernández, R., and Kruger, J. (2014). Empirical analysis of aggregation methods for collective annotation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1533–1542, Dublin, Ireland. Dublin City University and Association for Computational Linguistics. Cited on page 149.

Quan, C. and Ren, F. (2009). Construction of a blog emotion corpus for Chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1446–1454, Singapore. Association for Computational Linguistics. Cited on page 5.

Rabinovich, E., Patel, R. N., Mirkin, S., Specia, L., and Wintner, S. (2017). Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics. Cited on page 159.

Ramsey, F. P. (1927). Symposium: Facts and propositions. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 7:153–170. Cited on page 61.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics. Cited on pages 51 and 52.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., and Choi, Y. (2018). Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics. Cited on page 101.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics. Cited on pages 58 and 101.

Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., and Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, 37(5):420–445. Cited on page 127.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics. Cited on page 181.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., and Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3806–3813, Istanbul, Turkey. European Language Resources Association (ELRA). Cited on pages 39 and 45.

Romanov, A., Rumshisky, A., Rogers, A., and Donahue, D. (2019). Adversarial decomposition of text representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 51 and 52.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology*. Cited on page 86.

Roseman, I. J. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10(3):241–278. Cited on page 86.

Roseman, I. J. (2001). A model of appraisal in the emotion system. *Appraisal processes in emotion: Theory, methods, research*, pages 68–91. Cited on page 86.

Roseman, I. J., Spindel, M. S., and Jose, P. E. (1990). Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. *Journal of Personality and Social Psychology*, 59(5):899–915. Cited on page 86.

Roth, M. and Lapata, M. (2015). Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460. Cited on page 65.

Roth, M. and Lapata, M. (2016). Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting* *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202, Berlin, Germany. Association for Computational Linguistics. Cited on page 65.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121. Cited on page 46.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 41.

Ruppenhofer, J. (2018). The treatment of emotion vocabulary in framenet: Past, present and future developments. In Ziem, A., Inderelst, L., and Wulf, D., editors, *Frames interdisziplinär: Modelle, Anwendungsfelder, Methoden*, pages 95–122. Düsseldorf University Press. Cited on pages 66, 190, and 217.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., and Scheffczyk, J. (2016). Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute. Cited on page 63.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178. Cited on page 20.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145–172. Cited on page 32.

Russell, J. A. (2012). From a psychological constructionist perspective. In *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell*, volume 7, pages 79–118. John Benjamins Amsterdam, the Netherlands. Cited on pages 19 and 20.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294. Cited on pages 17 and 112.

Saadany, H. and Orasan, C. (2020). Is it great or terrible? preserving sentiment in neural machine translation of Arabic reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages

24–37, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on page 159.

Sabbatino, V., Troiano, E., Schweitzer, A., and Klinger, R. (2022). “splink” is happy and “phrouth” is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 37–50, Dublin, Ireland. Association for Computational Linguistics.

Sailunaz, K., Dhaliwal, M., Rokne, J., and Alhajj, R. (2018). Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26. Cited on page 44.

Salameh, M., Mohammad, S., and Kiritchenko, S. (2015). Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics. Cited on page 159.

Samonte, M. J. C., Punzalan, H. I. B., Santiago, R. J. P. G., and Linchangco, P. J. L. (2017). Emotion detection in blog posts using keyword spotting and semantic analysis. In *Proceedings of the 3rd International Conference on Communication and Information Processing*, pages 6–13. Cited on page 38.

Sander, D., Grandjean, D., and Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural networks*, 18(4):317–352. Cited on pages xi, 24, and 80.

Sarawagi, S. and Cohen, W. W. (2004). Semi-markov conditional random fields for information extraction. *Advances in neural information processing systems*, 17:1185–1192. Cited on page 197.

Scarantino, A. (2016). The philosophy of emotions and its impact on affective science. *Handbook of emotions*, 4:3–48. Cited on pages 3, 17, 20, 21, 26, and 30.

Scheff, T. J. (1973). Intersubjectivity and emotion. *American Behavioral Scientist*, 16(4):501–511. Cited on page 1.

Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of personality & social psychology*, pages 37–63. Cited on pages 21 and 219.

Scherer, K. R. (2000). Psychological models of emotion. In *The neuropsychology of emotion.*, Series in affective science., pages 137–162. Oxford University Press, New York, NY, US. Cited on page 16.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729. Cited on pages 23, 26, 31, 55, 72, and 94.

Scherer, K. R. (2011). On the rationality of emotions: or, when are emotions rational? *Social Science Information*, 50(3-4):330–350. Cited on page 122.

Scherer, K. R., Bänziger, T., and Roesch, E. (2010). *A Blueprint for Affective Computing: A sourcebook and manual.* Oxford University Press. Cited on pages 22, 25, and 80.

Scherer, K. R. and Fontaine, J. J. (2013). Driving the emotion process: The appraisal component. In Fontaine, J. J. R., Scherer, K. R., and Soriano, C., editors, *Series in affective science. Components of emotional meaning: A sourcebook*, chapter 12, pages 266–290. Oxford University Press, Oxford. Cited on pages xi, 23, 24, 80, 81, 83, and 84.

Scherer, K. R. and Grandjean, D. (2008). Facial expressions allow inference of both emotions and their components. *Cognition and Emotion*, 22(5):789–801. Cited on page 76.

Scherer, K. R., Schorr, A., and Johnstone, T. (2001a). *Appraisal considered as a process of multi-level sequential checking*, volume 92. Oxford University Press. Cited on page 86.

Scherer, K. R., Schorr, A., and Johnstone, T. (2001b). *Appraisal processes in emotion: Theory, methods, research.* Oxford University Press. Cited on page 22.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310. Cited on pages 47, 75, and 166.

Scherer, K. R. and Wallbott, H. G. (1997). The ISEAR questionnaire and codebook. Geneva Emotion Research Group. Cited on pages xv, 80, 81, 82, 86, 92, 97, 141, 195, 238, and 242.

Schmid, P. C. and Schmid Mast, M. (2010). Mood effects on emotion recognition. *Motivation and Emotion*, 34(3):288–292. Cited on page 27.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics. Cited on pages 31, 35, 42, 44, 124, 141, and 195.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681. Cited on page 40.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., and Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and individual differences*, 25(2):167–177. Cited on page 28.

Shaikh, M. A. M., Prendinger, H., and Ishizuka, M. (2009). A linguistic interpretation of the occ emotion model for affect sensing from text. *Affective Information Processing*, pages 45–73. Cited on pages 37, 46, 75, and 219.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, pages 6830–6841. Curran Associates, Inc. Cited on page 52.

Shivhare, S. N., Garg, S., and Mishra, A. (2015). Emotionfinder: Detecting emotion from blogs and textual documents. In *International Conference on Computing, Communication & Automation*, pages 52–57. IEEE. Cited on page 37.

Shuster, K., Humeau, S., Hu, H., Bordes, A., and Weston, J. (2019). Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12516–12526. Cited on page 58.

Si, C., Wu, K., Aw, A. T., and Kan, M.-Y. (2019). Sentiment aware neural machine translation. In *Proceedings of the 6th Workshop on*

Asian Translation, pages 200–206, Hong Kong, China. Association for Computational Linguistics. Cited on page 160.

Sidorov, M., Brester, C., Minker, W., and Semenkin, E. (2014). Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3481–3485, Reykjavik, Iceland. European Language Resources Association (ELRA). Cited on page 30.

Smith, C. A. and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838. Cited on pages xv, 23, 25, 72, 80, 81, 82, 84, 85, 86, 95, 208, 238, and 242.

Smith, E. M., Gonzalez-Rico, D., Dinan, E., and Boureau, Y. (2019). Zero-shot fine-grained style transfer: Leveraging distributed continuous style representations to transfer to unseen styles. *CoRR*, abs/1911.03914. Cited on pages 52, 57, 58, 162, and 231.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American economic review*, 72(5):923–955. Cited on page 121.

Sommerauer, P., Fokkens, A., and Vossen, P. (2020). Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics. Cited on page 149.

Song, Z., Zheng, X., Liu, L., Xu, M., and Huang, X. (2019). Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics. Cited on page 52.

Sonnemans, J. and Frijda, N. H. (1994). The structure of subjective emotional intensity. *Cognition & Emotion*, 8(4):329–350. Cited on page 183.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):71–101. Cited on page 144.

Štajner, S. (2021). Exploring reliability of gold labels for emotion detection in Twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1350–1359, Held Online. INCOMA Ltd. Cited on pages 6, 35, and 111.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2009). The occ model revisited. Online: `https://people.idsia.ch/~steunebrink/Publications/KI09_OCC_revisited.pdf`. Cited on pages xi and 21.

Stranisci, M. A., Frenda, S., Ceccaldi, E., Basile, V., Damiano, R., and Patti, V. (2022). APPReddit: a corpus of reddit posts annotated for appraisal. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association. Cited on page 120.

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics. Cited on pages 34, 44, and 139.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA. ACM. Cited on pages 36, 43, 44, and 140.

Strapparava, C. and Valitutti, A. (2004). WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA). Cited on pages 36, 43, and 188.

Sudhakar, A., Upadhyay, B., and Maheswaran, A. (2019). "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics. Cited on page 51.

Sun, K., Luo, X., and Luo, M. Y. (2022). A survey of pretrained language models. In *International Conference on Knowledge Science, Engineering and Management*, pages 442–456. Springer. Cited on page 40.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics. Cited on page 176.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc. Cited on page 41.

Sutton, C. and McCallum, A. (2005). Joint parsing and semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 225–228, Ann Arbor, Michigan. Association for Computational Linguistics. Cited on page 64.

Swayamdipta, S., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Greedy, joint syntactic-semantic parsing with stack LSTMs. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 187–197, Berlin, Germany. Association for Computational Linguistics. Cited on page 65.

Swayamdipta, S., Thomson, S., Dyer, C., and Smith, N. A. (2017). Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*. Cited on pages xv, 65, and 197.

Täckström, O., Ganchev, K., and Das, D. (2015). Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41. Cited on page 65.

Tarski, A. (1983). *Logic, semantics, metamathematics: papers from 1923 to 1938*. Hackett Publishing Company. Cited on page 60.

Tebbifakhr, A., Bentivogli, L., Negri, M., and Turchi, M. (2019). Machine translation for machines: the sentiment classification use case. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1368–1374, Hong Kong, China. Association for Computational Linguistics. Cited on page 179.

Terracciano, A., Merritt, M., Zonderman, A. B., and Evans, M. K. (2003). Personality traits and sex differences in emotions recognition among african americans and caucasians. *Annals of the New York Academy of Sciences*, 1000:309. Cited on page 126.

Tikhonov, A. and Yamshchikov, I. P. (2018). What is wrong with style transfer for texts? *arXiv preprint arXiv:1808.04365*. Cited on page 183.

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*, volume I. Springer publishing company. Cited on page 17.

Tooby, J. and Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In M. Lewis, J. H.-J. and Barrett, L. F., editors, *Handbook of emotions*, pages 114–137. Guilford Press, New York. Cited on page 17.

Toprak, C., Jakob, N., and Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics. Cited on page 75.

Toshevska, M. and Gievska, S. (2021). A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*. Cited on page 51.

Toutanova, K., Haghighi, A., and Manning, C. D. (2008). A global joint model for semantic role labeling. *Computational Linguistics*, 34(2):161–191. Cited on page 64.

Tracy, J. L. and Robins, R. W. (2006). Appraisal antecedents of shame and guilt: Support for a theoretical model. *Personality and social psychology bulletin*, 32(10):1339–1351. Cited on pages 26, 90, 96, 112, and 174.

Trauffer, N. M., Widen, S. C., and Russell, J. A. (2013). Education and the attribution of emotion to facial expressions. *Psihologijske teme*, 22(2):237–247. Cited on page 126.

Troiano, E., Klinger, R., and Padó, S. (2020). Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Troiano, E., Klinger, R., and Padó, S. (2023a). On the relationship between frames and emotionality in text. *Northern European Journal of Language Technology*, 9(1).

Troiano, E., Oberländer, L., and Klinger, R. (2023b). Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72. Cited on pages 116 and 117.

Troiano, E., Oberländer, L. A. M., Wegge, M., and Klinger, R. (2022a). x-enVENT: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association. Cited on pages 101 and 267.

Troiano, E., Padó, S., and Klinger, R. (2019). Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Troiano, E., Padó, S., and Klinger, R. (2021). Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.

Troiano, E., Velutharambath, A., and Klinger, R. (2022b). From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*, pages 1–60. Cited on page 53.

Udochukwu, O. and He, Y. (2015). A rule-based approach to implicit emotion detection in text. In Biemann, C., Handschuh, S., Freitas, A., Meziane, F., and Métais, E., editors, *Natural Language Processing and Information Systems*, pages 197–203, Cham. Springer International Publishing. Cited on pages 46 and 219.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., and Poesio, M. (2021). SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics. Cited on page 124.

Ushio, A., Liberatore, F., and Camacho-Collados, J. (2021). Back to the basics: A quantitative analysis of statistical and graph-based term weighting schemes for keyword extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8089–8103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 198.

Van Kleef, G. A., Van Doorn, E. A., Heerdink, M. W., and Koning, L. F. (2011). Emotion is for influence. *European Review of Social Psychology*, 22(1):114–163. Cited on page 1.

Vanecek, E. and Dressler, W. (1975). Bericht uber psycholinguistische experimente zur sprechvariation. *Weiner Linguistische Gazette*, 9:17–38. Cited on page 55.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Cited on pages 40 and 166.

Vo, M., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., and Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, 41(2):534–538. Cited on page 72.

Volkova, E. P., Mohler, B., Meurers, D., Gerdemann, D., and Bülthoff, H. H. (2010). Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106, Los Angeles, CA. Association for Computational Linguistics. Cited on pages 35 and 124.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2012). Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 587–592. IEEE. Cited on page 45.

Wang, X. and Zheng, Q. (2013). Text emotion classification research based on improved latent semantic analysis algorithm. In *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pages 210–213. Atlantis Press. Cited on page 38.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207. Cited on page 45.

Wassmann, C. (2017). Forgotten origins, occluded meanings: Translation of emotion terms. *Emotion Review*, 9(2):163–171. Cited on page 159.

Wauthier, F. L. and Jordan, M. (2011). Bayesian bias mitigation for crowdsourcing. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc. Cited on pages 33 and 277.

Wei, P., Zhao, J., and Mao, W. (2020). Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online. Association for Computational Linguistics. Cited on page 189.

Wen, Z., Cao, J., Yang, R., and Wang, S. (2020). Decode with template: Content preserving sentiment transfer. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4671–4679, Marseille, France. European Language Resources Association. Cited on page 51.

Widen, S. C. (2013). Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emotion Review*, 5(1):72–77. Cited on page 126.

Wierzbicka, A. (1995). Emotion and facial expression: A semantic perspective. *Culture & Psychology*, 1(2):227–258. Cited on page 43.

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge university press. Cited on page 20.

Wierzbicka, A. (2013). *Imprisoned in English: The hazards of English as a default language*. Oxford University Press. Cited on page 159.

Wieting, J. and Gimpel, K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics. Cited on page 158.

Williams, J., Kleinegesse, S., Comanescu, R., and Radu, O. (2018). Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics. Cited on page 29.

Wilson, T. (2008). Annotating subjective content in meetings. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Cited on page 75.

Wispé, L. (1986). The distinction between sympathy and empathy: To call forth a concept, a word is needed. *Journal of personality and social psychology*, 50(2):314. Cited on page 27.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul. Cited on page 60.

Wolfgang, A. and Cohen, M. (1988). Sensitivity of Canadians, Latin Americans, Ethiopians, and Israelis to interracial facial expressions of emotions. *International Journal of Intercultural Relations*, 12(2):139–151. Cited on page 126.

Wood, I., McCrae, J. P., Andryushechkin, V., and Buitelaar, P. (2018). A comparison of emotion annotation schemes and a new annotated data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Cited on page 5.

Wu, C., Ren, X., Luo, F., and Sun, X. (2019a). A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883, Florence, Italy. Association for Computational Linguistics. Cited on page 52.

Wu, X., Xu, K., and Hall, P. (2017). A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology*, 22(6):660–674. Cited on page 48.

Wu, X., Zhang, T., Zang, L., Han, J., and Hu, S. (2019b). Mask and infill: Applying masked language model for sentiment transfer. In

Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, (IJCAI-19), pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization. Cited on page 51.

Xia, R. and Ding, Z. (2019). Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics. Cited on page 71.

Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., and Li, W. (2018). Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics. Cited on page 51.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415. Cited on page 159.

Xu, W., Ritter, A., Dolan, B., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee. Cited on page 51.

Yamashita, N., Inaba, R., Kuzuoka, H., and Ishida, T. (2009). Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 679–688, New York, NY, USA. Association for Computing Machinery. Cited on page 159.

Yamshchikov, I. P., Shibaev, V., Nagaev, A., Jost, J., and Tikhonov, A. (2019). Decomposing textual information for style transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 128–137, Hong Kong. Association for Computational Linguistics. Cited on page 182.

Yanchus, N. J. (2006). *Development and validation of a self-report cognitive appraisal scale*. PhD thesis, University of Georgia. Cited on page 81.

Yang, B. and Mitchell, T. (2017). A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Copenhagen, Denmark. Association for Computational Linguistics. Cited on page 65.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. R., and Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics. Cited on pages 34, 42, and 46.

Yuan, Z. and Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics. Cited on page 159.

Zaninello, A. and Birch, A. (2020). Multiword expression aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association. Cited on page 158.

Zhang, D., Chen, X., Xu, S., and Xu, B. (2020a). Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4429–4440, Barcelona, Spain (Online). International Committee on Computational Linguistics. Cited on page 3.

Zhang, L. and Liu, B. (2011a). Extracting resource terms for sentiment analysis. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1171–1179, Chiang Mai, Thailand. Asian Federation of Natural Language Processing. Cited on page 75.

Zhang, L. and Liu, B. (2011b). Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 575–580, Portland, Oregon, USA. Association for Computational Linguistics. Cited on page 75.

Zhang, Y., Ge, T., and Sun, X. (2020b). Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics. Cited on page 51.

Zhao, J. J., Kim, Y., Zhang, K., Rush, A. M., and LeCun, Y. (2018). Adversarially regularized autoencoders. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR. Cited on page 52.

Zhao, Z., Salesse, R. N., Qu, X., Marin, L., Gueugnon, M., and Bardy, B. G. (2020). Influence of perceived emotion and gender on social motor coordination. *British Journal of Psychology*, 111(3):536–555. Cited on page 1.

Zhou, J. and Bhat, S. (2021). Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 49.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China. Association for Computational Linguistics. Cited on page 64.

Zhou, X. and Wang, W. Y. (2018). MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics. Cited on page 52.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China. Cited on pages 40, 46, 116, and 277.

Zull, J. E. (2006). Key aspects of how the brain learns. *New Directions for Adult and Continuing Education*, 110:1–10. Cited on page 57.