Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master thesis

Investigating Topic Bias in Emotion Classification

Maximilian Wegge

Studiengang: M.Sc. Computational Linguistics

Prüfer*innen: Prof. Dr. Roman Klinger
Prof. Dr. Sebastian Padó
Betreuer: Prof. Dr. Roman Klinger

Beginn der Arbeit: 09.01.2023
Ende der Arbeit: 09.07.2023

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.

Statement of Authorship

This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

(Maximilian Wegge)

Abstract

In emotion classification, texts are assigned a conceptual emotion representation such as discrete labels or dimensions of cognitive appraisal. Emotion classifiers are typically not universally applicable, but base their classification decisions on characteristic features of a specific domain. When applied to a different domain, the lack of domain-specific knowledge results in classification errors. While this behavior is typically addressed as a cross-domain or cross-corpus phenomenon, the potentially misguiding factors within one corpus have not yet been studied to the same degree. I propose an investigation of topics in emotion datasets to assess their influence on the classification decisions in emotion and appraisal classification. My contribution is threefold: First, I conduct an analysis of how topics and emotions are distributed in emotion datasets. Second, I investigate whether state-of-the-art emotion classification systems are prone to adopting the topic distribution in the training data as topic bias. Third, I evaluate debiasing methods for topic bias in the context of emotion classification. The results indicate that topic bias is introduced to emotion datasets through the applied sampling method. The topic bias within commonly used datasets in the field appears to be, except for one exception, negligible. However, if bias is present in the data, it is adopted by the resulting classifiers. In order to mitigate such bias, I investigate a naive word removal approach as well as gradient reversal, which is found to work best for bis mitigation.

Contents

1 Introduction

Emotions are essential to how we experience the world around us: They affect our decision-making, influence the way we react to particular situations and enable us to express and communicate our feelings. The question of "What is an emotion?" (James, 1884) has therefore attracted the interest of various research disciplines including psychology, philosophy, neuroscience and biology, and – as an integral part of communication and language – linguistics. Linguistic research on emotions is leveraged by natural language processing, which allows to study how emotions manifest in language on a large scale. Since a significant amount of modern communication and information is text-based, computational emotion analysis is typically performed on written texts. Approaches in the field cover various domains ranging from political debates (Mohammad et al., 2014) to dialog (Li et al., 2017) and literary texts (Mohammad, 2011), and enable further use cases such as analyzing social media users' emotions in response to the COVID-19 pandemic (Zhan et al., 2022), identifying abusive language using emotional cues (Safi Samghabadi et al., 2020) or developing empathetic dialog agents, e.g., for emotional support (Liu et al., 2021). Across domains and down-stream tasks, emotion analysis is usually formulated as the task of emotion classification, i.e., assigning emotions to textual units such as news headlines, dialog or social media and blog posts.

Although emotions play a central role in various research disciplines, there is no unified theory of what constitutes an emotion. In consequence, there are several theoretical models of emotion that have been adopted for emotion classification. Among the most influential are theories of basic emotions, grounded in evolutionary theory. Based on fundamental facial expressions, Ekman (1992) defines the set of basic emotions as *anger*, *fear*, *joy*, *sadness*, *disgust* and *surprise*. The basic emotion model proposed by Plutchik (2001) considers *trust* and *anticipation* in addition to the six emotions defined by Ekman, and further accounts for emotion intensity and for the relations between emotions.

Some approaches to emotion analysis instead adopt dimensional models of emotion (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017a). Rather than modeling emotions as discrete classes, dimensional models map emotion labels to coordinates in a continuous vector space. Following the model of *core-affect* by Russell (1980), such vector space could

be defined along the dimensions of *valence* (the pleasantness of the event) and *arousal*.

Recently, theories of cognitive appraisal were adopted for computational emotion analysis as well (Troiano et al., 2023; Hofmann et al., 2020; Stranisci et al., 2022). Appraisal theories hone in on the cognitive component of emotions (Scherer, 2005), modelling emotions as the result of the emoter's cognitive appraisal of the stimulus event. Multiple appraisal dimensions have been proposed. Smith and Ellsworth (1985), for example, consider a total of six dimensions by which an event is appraised: *pleasantness*, *control*, *certainty*, *attention*, *effort* and *situational control*. Emotions are then characterized by the combination of appraisal dimensions that hold in the given situation (e.g., a low appraisal of self-responsibility, self-control and certainty in combination with a strong appraisal of unpleasantness might indicate *fear*).

Independent of which emotion theory is adopted, emotion classification from text faces a fundamental challenge: In contrast to emotion recognition systems that use other modalities, such as speech or vision, text-based systems are limited to textual features. This is especially challenging whenever explicit cue words are absent, since the classification decision is then resorted to highly ambiguous context features. In order to cope with this ambiguity and still enable capable emotion classification, computational approaches to emotion analysis are typically geared towards specific domains. Domain-specific features include, i.a., the writing style (e.g., news articles differ significantly from social media posts in terms of sentence length, choice of words, etc.) or domain-specific terms, as some terms might only appear in one specific domain or have a different emotional connotation across domains.

Therefore, emotion classifiers as well as the datasets needed for developing them are specific to their respective domain and task. In consequence, the classifiers' ability to generalize is other domains is typically very limited, resulting in weaker performance when applied to a different dataset than the one it was developed on.

This bias towards specific domains in emotion classification is well acknowledged even has its own field of research associated to it, namely domain adaptation. However, domain adaptation approaches the issue of dataset bias as a cross-domain, i.e., cross-corpus phenomenon. What has not yet been explored is whether emotion datasets are also

inherently biased.

Precisely, this work investigates topic-specific bias, which is inherently related to emotions. Emotions are associated with stimulus events, since they – depending on the underlying emotion theory – emerge from or consist of the emoter's immediate reaction to the trigger event. As such, some emotions are prototypical of certain events. For example, "birthday party" is generally likely to be associated with a positive emotion such as *joy*, and "funeral", on the other hand, with a negative emotion (e.g., *sadness*). While the emotional associations of topics such as "birthday party" or "funeral" are very likely to hold universally across different domains and text styles, there might be a biased representation of other topics and associated emotions in the training data. For instance, if the topic of "relationship" appears mainly in a negative context in the training data (e.g., if part of the data is subsampled from a relationship counseling forum[1]), the emotion classifier might treat topic features as cues for negative emotions, instead of relying on other textual features actually related to emotions. As a result, texts containing positive emotions in the context of relationships might then be misclassified. Topic bias thus affects texts that are sampled from the training domain, as well as – to an even greater extent – texts from a different domain.

While the issue of topic bias in emotion datasets has not been explicitly examined yet in the context of emotion classification, there are investigations in emotion and sentiment analysis that touch on bias, most notably in the context of social bias (Kiritchenko and Mohammad, 2018; Câmara et al., 2022). Social bias denotes bias towards certain social groups (defined by gender, ethnicity or age) that are over- or underrepresented within the data, or that appear only in specific (negative) contexts (Spliethöver and Wachsmuth, 2020b). Zad et al. (2021) identify erroneous entries in the influential NRC emotion lexicon (Mohammad and Turney, 2013) that they ascribe to the NRC's lack of domain-specificity, missing part of speech indications as well as to "simple errors". While some of these errors

[1]For example, the subreddit r/relationships provides a platform for users to seek and provide relationship advice. Therefore, the users' descriptions of their relationships mainly feature negative aspects (e.g., "Throughout our relationship she's been complaining non-stop"). Sampling part of the training data from this subreddit would likely result in an over-representation of "relationship"-related texts associated with negative emotions.

are seemingly based on social bias (e.g., "mosque" associated with *anger*), others seem to express topic bias (e.g., "stone" associated with *anger*). However, Zad et al. do not further discuss the different types of biases.

Topic bias is explicitly addressed in disciplines other than emotion analysis (e.g., hate-speech detection, Wich et al., 2020; Wiegand et al., 2019; or quality assessment, Ferschke et al., 2013), alongside related biases such as authorship bias. Previous work in these fields has also proposed various methods for mitigating bias, either at the stage of data collection (Wiegand et al., 2019; Barikeri et al., 2021) or at modeling-level (Elazar and Goldberg, 2018). This thesis introduces the investigation of topic bias to emotion classification, bridging the gap between computational emotion analysis and research on bias in related disciplines. The investigation of topic bias is conducted along the following research questions:

Are emotion datasets biased towards topics? Bias might be introduced to a computational system at multiple points during its development. Arguably the most critical component in that context is the data used for training and tuning a computational model. I leverage a topic modeling approach to conduct an analysis of established and widely used corpora for emotion and appraisal classification, in order to assess whether they comprise an underlying bias towards specific topics.

Is emotion classification influenced by topics? The central contribution of this thesis is to assess whether emotion classifiers are prone to towards topic bias in text, e.g., by adopting such bias from the training data. Since emotions have not been investigated in the context of topics yet, this thesis provides a novel perspective on emotion analysis and, ultimately, contributes towards a better understanding of how emotions are conveyed in text. The proposed account of topic bias in emotion analysis explicitly extends to computational classification of appraisals as well.

Can the influence of topics on emotion classification be mitigated? Under the assumption that emotion classification is indeed biased towards topics, it has to be assessed

whether this bias can be mitigated. To this end, I propose two debiasing approaches, one adversarial, the other based on removing topic-relevant lexical features, thus contributing towards more robust emotion classification.

In order to provide meaningful results which do not depend on the individual characteristics of a single dataset, all experiments are conducted using a total of six different emotion corpora: ISEAR (Scherer and Wallbott, 1994), SSEC (Schuff et al., 2017), TALES (Alm et al., 2005), CROWD-ENVENT (Troiano et al., 2023), APPREDDIT (Stranisci et al., 2022) and ENISEAR (Hofmann et al., 2020).

These corpora were selected based on three criteria: They cover a **(i) variety of domains**, including literature, social media posts and event descriptions. In addition to their respective domain, the corpora differ in their **(ii) choice of emotion model**. Most notably, three out of six corpora are annotated for appraisal dimensions, namely APPREDDIT, ENISEAR and CROWD-ENVENT (CROWD-ENVENT is annotated with both discrete emotion labels and appraisal dimensions.). Although now established, appraisal classification is a comparatively recent task in emotion analysis. By investigating topic bias explicitly in the context of both appraisals and (basic) emotions, this thesis contributes towards a deeper understanding of how appraisals are conveyed in text and how they relate to emotions. All corpora are further **(iii) well-established** resources for the development of emotion and appraisal classification systems (Bostan and Klinger, 2018; Dong and Zeng, 2022). The findings on topic bias presented in this thesis can thus be applied to a variety of previous approaches.

The following section 2 provides an overview over the theoretical and practical research that this work builds upon. Section 3 discusses the concrete methods used for investigating and mitigating topic bias, and presents the data used in the experiments in more detail (3.4). The results are presented and discussed in section 4; section 6 concludes the thesis.

2 Background and Related Work

The investigations conducted in this thesis combine research from NLP, emotion theory and bias. This section provides the theoretical background for the subsequent experiments,

consisting of influential emotion theories in psychology (2.1), emotions in the context of NLP (2.2), the challenges in computational emotion classification (2.3; 2.3.2) and bias in computational resources (2.4).

2.1 Emotion Theories

Despite their fundamental role in human affective live, emotions are a comparatively recent subject in psychological research. The term "emotion" was first introduced in the mid-nineteenth century, replacing the previous perspective of passions and affects (Cooke, 1838; Ramsay, 1848). Although emotions are now an established research field in psychology and researchers acknowledge their relevance (Dixon, 2012; Scarantino and de Sousa, 2021; Smith and Lazarus, 1990), there is still no consensus on the exact definition of emotions (Izard, 2010). What is largely accepted, however, is the understanding of emotions as complex events comprised of several components. Different emotion theories diverge on the question of which components contribute (most) towards the experience of emotion: physiological (bodily reactions, such as sweating or increased heart rate), phenomenological (the subjective feeling), expressive (facial or vocal expressions), behavioral (motor reactions, e.g., running towards or from something) or cognitive (appraisal of an event) aspects (c.f. Scarantino and de Sousa, 2021). Further, emotion theories differ in how emotions are distinguished from one another (and from non-emotions), what their primary purpose is and how they are elicited. Based on how different theories approach these questions, they can be divided into three overall research traditions: The *Feeling Tradition*, *Motivational Tradition* and *Evaluative Tradition* (Scarantino and de Sousa, 2021).

2.1.1 Feeling Tradition

The Feeling Tradition puts the most intuitive aspect of emotions, the subjective experience, at center. Its most prominent theory by James (1884) defines emotions as the feeling, i.e., experience of physiological changes in response to an emotion-eliciting event. For example, if a person with severe fear of heights is standing besides a steep drop, their palms start sweating, their heart and breathing rate increase, etc. Experiencing these bodily changes (and being aware of them) thus makes an emotion (e.g., fear).

It is important to note that, according to James (1884), emotions emerge from (the perception of) the bodily reactions to a stimulus, and not the other way around. By this, James denies emotions any involvement in the process of eliciting the bodily reactions. In consequence, the theory cannot explain what initially causes the bodily reactions.

2.1.2 Motivational Tradition

Opposed to the Feeling Tradition, which does not grant any motivational relevance to emotions, the Motivation Tradition understands emotions specifically as motivational states, i.e., as "internal cause of behaviors aimed at satisfying a goal" (Scarantino and de Sousa, 2021). This perspective is prominently featured in theories of Basic Emotions. In this context, emotions can be understood as affect programs, which, when activated, motivate appropriate behavior, cause certain (facial) expressions, recall relevant memories, etc. (Ekman and Cordaro, 2011). Basic emotions are thus grounded in evolutionary theory, as they serve in dealing with the most "fundamental life tasks" (Ekman, 1999; p. 46), such as survival, coping with loss or achieving a goal. For example, the experience of *fear* might provide the motivation to fight off a predator, or, alternatively, flee to safety. Concerning the question of differentiation, Ekman proposes several criteria that distinguish basic emotions, including, i.a., distinctive physiology, brief duration and correspondence in other animals (the initial nine criteria were later extended to eleven in Ekman, 1999). Among these, the criterion of distinctive universal signals, i.e., facial expressions, is the most significant. Ekman (1992) finds evidence for six basic emotions (*anger*, *fear*, *joy*, *sadness*, *disgust* and *surprise*), each of which correspond to a distinct facial expression. The studies conducted by Ekman indicate that these facial expressions are universal, i.e., are recognized across cultures.

Another theory to the Motivational Tradition was proposed by Plutchik (1982). Honing in on the evolutionary purpose of emotions, Plutchik assigns each basic emotion a purpose for survival in the form of prototypical behavioral patterns: rejection, protection, destruction, reproduction, reintegration, orientation, exploration and incorporation/ affiliation (Plutchik, 1982; p. 537). According to these patterns, the set of basic emotions consists of *disgust*, *fear*, *anger*, *joy*, *sadness*, *surprise*, *anticipation* and *trust* (introduced in Plutchik,

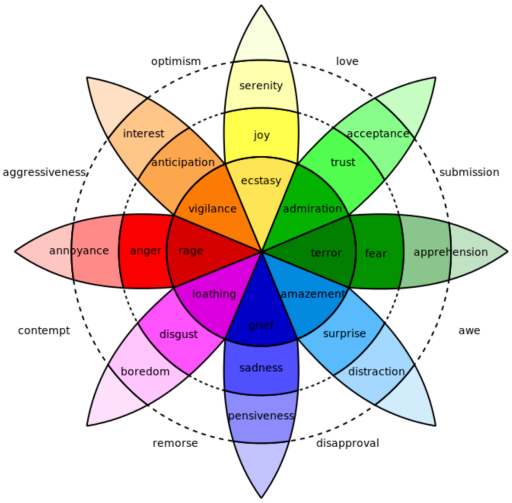

Figure 1: Wheel of Emotions as proposed by Plutchik (2001). combining basic emotions with the dimensions of emotion intensity, similarity and polarity.

2001; initially proposed as *acceptance*). Plutchik further extends the model of basic emotions beyond discrete classes by incorporating the dimensions of intensity, similarity and polarity. In the resulting wheel of emotions (1), similar emotions are depicted adjacent to each other (e.g., *fear* and *surprise*), while polar emotions are placed at opposing sides of the wheel (emotions on the outer circle are more similar to each other than the ones on the inner circle). Emotion intensity increases from the outer to the inner circle (e.g., *rage* is more intense than *annoyance*). Analogous to blending colors on a color wheel, the model accounts for "mixed" emotions, i.e., emotions that are neither considered basic nor an intensity-variation of it, but a blend of adjacent emotion pairs blend (e.g., *love*, consisting of *trust* and *joy*).

2.1.3 Evaluative Tradition

The Evaluative Tradition emerged at around 1960 in the context of the rise of cognitivism in psychology. Theories of the Evaluative Tradition oppose the definition of emotions as feelings. Instead, appraisal theories define emotions through the cognitive evaluation of the stimulus. In the Component Process Model by Scherer (2001), emotions are defined as

"an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event" (Scherer, 2005; reported from Scarantino and de Sousa, 2021). The theory thus establishes cognitive appraisal as one of the components that comprise emotions[2], alongside the more canonical components of bodily symptoms (neurophysiological component), action tendencies (motivational component), facial expression (motor expression component) and emotional experience (subjective feeling component).

Scherer (2001) divides the appraisal process into four sequential stages, each dedicated to evaluating a different aspect of the stimulus, increasing in complexity: the relevance of the stimulus, its implications, the emoter's potential to cope with the consequences and the significance for social and personal norms. The stimulus is evaluated based on sixteen distinct appraisal dimensions, each mapped to the respective stage of the evaluation process. For example, the relevance of the stimulus is appraised along the dimensions of *pleasantness* (how pleasant is the event?), *novelty* (how novel is the event?) and *goal relevance* (how relevant is the event in relation to my personal goals?). As the event is constantly re-evaluated, the appraisal of certain dimensions might change, potentially resulting in a change of the overall emotion as well. Appraisal theories further account for the case that the same event elicits different emotions in different emoters: Since the appraisal dimensions are specific to the individual emoter (e.g., *is the event in line with my personal standards?*), the emotion comprised of these individual appraisals is also specific to the emoter.

There are multiple theories of appraisal (Smith and Ellsworth, 1985; Lazarus, 1991; Scherer and Fontaine, 2013) that diverge in their choice of appraisal dimensions. Smith and Ellsworth (1985), for example, consider the six appraisal dimensions of *pleasantness*, *self responsibility* (to what degree is emoter is responsible for the event), *self control* (the level of control the emoter appraises to have over the situation), *effort* (the degree of expected effort), *attention* (the level of attention the emoter has to pay to the event) and *certainty* (the degree of certainty about what is going to happen) ,which allow them to define 15 individual emotions based on combinations of appraisal values. Further, there is

[2]each of the organismic subsystems that Scherer's definition is based on is instantiated in the resepective emotion component.

a distinction between of constitutive and causal appraisal theories. Causal theories (e.g., Scherer, 2001) understand appraisals to cause the emotion, while in constitutive theories (e.g., Smith and Ellsworth, 1985), emotions consist of appraisals (cf. Troiano et al., 2023; Scarantino and de Sousa, 2021).

2.1.4 Affect and Sentiment

In psychology, two terms that occur repeatedly in the psychological context of emotions – sometimes even synonymously – are "affect" and "sentiment" (alongside further terms, such as "opinion", "mood" and "feeling"; refer to Munezero et al., 2014 for an overview). In order to conclude the theoretical background on emotion theories, these two terms will be put into perspective regarding the concept of emotion.

While emotions are typically defined based on physiological, phenomenological, expressive, behavioral and cognitive components, sentiments are related to social aspects, understood as "socially constructed patterns of sensations, expressive gestures, and cultural meanings organized around a relationship to a social object, usually another person" (Gordon, 1981; reported from Munezero et al., 2014). Sentiments are also considered to be of longer duration than the immediate reactions to specific stimuli that are emotions. For example, one could maintain the sentiment of *love* towards someone over an extended period of time without constantly experiencing the emotion of *love*. The same applies for *friendship* or *hate*, and also more acute sentiments such as pride or grief (Munezero et al., 2014) [3]. Sentiment is an extensive field of research within psychology; however, since out of the scope of this thesis, it will not be further discussed here.

Affect acts as an umbrella term for emotion and sentiment, as well as for all related concepts (e.g., feelings, opinions, etc.), and can be interpreted as "a predecessor to feelings and emotions" (Munezero et al., 2014; p. 104). As such, affect is the most abstract concept of those discussed here (illustrated by Batson et al., 1992: "affect is present in the yelp of a dog and in the coo or cry of an infant") It cannot be classified by means of discrete labels, but rather along dimensions such as polarity (positive-negative), intensity or activation (Thoits, 1989).

[3] Note that the understanding of sentiment differs across psychology and NLP (cf. 2.2).

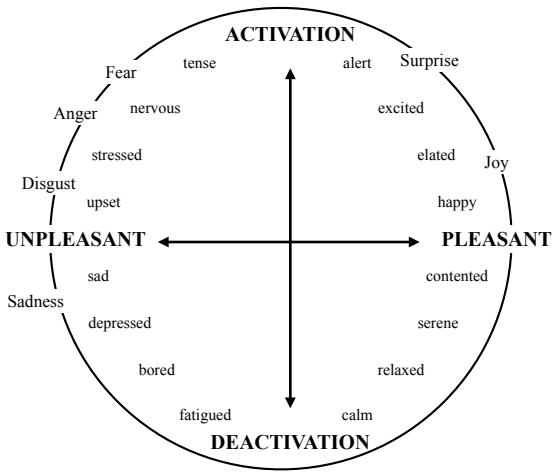

Figure 2: Circumplex model of Affect (Russell, 1980). Core-affect is modeled as a two-dimensional vector space of *valence* and *arousal* (adapted from Feldman Barrett and Russell, 1998).

In this sense, Russell introduces the concept of *core-affect*, a "neurophysiological state consciously accessible as the simplest raw (nonreflective) feelings evident in moods and emotions" (Russell, 2003; p. 148). Core-affect is defined with respect to *valence* and *arousal*, where *valence* indicates the degree of pleasantness. The circumplex model of affect (2) links affect with emotion theory by mapping emotion labels to their respective values within the vector-space of *valence* and *arousal*. The circumplex model thus represents a constructionist theory of emotions, as it assumes that emotions arise spontaneously from the underlying, ubiquitous state of core-affect (Scarantino and de Sousa, 2021; cf.). For example, *anger* and *sadness* are both characterized by low values of pleasantness, but *anger* scores higher on arousal than *sadness*.

2.2 Emotions in Language

Emotions take on a central role in conveying information in social interactions of various kinds. The communicative purpose of emotions comprises very fundamental information

needs (e.g., the presence of an immediate threat by expressing *fear*) as well as complex social strategies, as exemplified by the *audience-effect*. The effect describes the phenomenon that the same stimulus might elicit different displays of emotion, depending on the emoter's audience (e.g., in a competition, the expression of intense *joy* when executing a good move in front of the opponent, opposed to the lack of such expression when executing that same move, but facing away from the opponent; cf. Griffiths and Scarantino, 2005). By definition, the most prominent means of communicating emotions is their expressive component, i.e., the display of facial expressions (and, to some extent, also the physiological component comprising bodily features). The question that arises from this, is how emotions are conveyed beyond these non-verbal means, specifically in language. For speech, the answer is relatively straightforward, as the expressive component of emotions expands to the production of sounds as well [4], which involves phonetic and prosodic features such as pitch, intensity, etc. In written language, however, these features are absent. Nonetheless, texts such as "I am just so happy today!", "My cat passed away." or "I told you we should have left earlier!", appear to convey some sort of emotional information (albeit each in a different way). Scarantino and de Sousa (2021) even identify the compatibility of emotion concepts and their linguistic usage as one central goal in the efforts towards defining emotions: "to secure ordinary language compatibility, traditional philosophers have relied on introspection, thought experiments, casual observation, gleaning of insights from literary texts and other artistic sources". However, the relation between emotions and their linguistic representations appears to be more complex than a one-to-one equivalence. In her account of a universal emotion language (1992, 1995), Wierzbicka argues that emotion words impose a language- and culture-specific classification on the respective emotion. For example, the English *anger* might be translated into another language, but the specific concept associated with *anger* is lost during translation and instead replaced by the meaning of the translation. In order to diminish this language-bias in emotion research, Wierzbicka proposes a universal language of "primitives", i.e., basic affect words that have equivalents across all languages (e.g., feel, want, good, bad). There is further evidence for a connection between emotion and language from neuroscience, as neuroimaging studies

[4] In Scherer's component process model, vocal expression is explicitly mentioned alongside facial expression as the motor expression component of emotions.

suggest that brain regions typically associated with language processing are also activated during emotional episodes (Lindquist et al., 2015; Lindquist, 2017).

Understanding how emotions manifest in language is therefore essential for both linguistic and psychological research. For psychology, it offers insight into the cognitive processes underlying emotions. Correspondingly, the benefit for linguistic research is a deeper understanding of how emotions are communicated in text.

2.3 Computational Emotion Analysis

Investigating emotions in texts requires analyzing a large number of samples across different domains, text types, emoters (i.e., in the context of text, authors) and languages. Since most of today's communication takes place online, this data is readily available, namely in the form of social media posts, forum debates or blog entries. As manual analyses of such large amounts of data is not feasible, automatic, i.e., computational systems for emotion analysis are required.

Computational approaches to emotion analysis typically adopt theories of basic emotions, following either Ekman (1999) or Plutchik (1982), by modeling emotions as either six (*anger*, *fear*, *joy*, *sadness*, *disgust*, *surprise*) or eight (*anger*, *fear*, *joy*, *sadness*, *disgust*, *surprise*, *anticipation*, *trust*) discrete classes, respectively. Some approaches also follow Russell's dimensional approach by modeling emotions within the vector space of *valence* and *arousal* (Yu et al., 2016 Preoţiuc-Pietro et al., 2016). Buechel and Hahn (2017a) additionally consider the dimension of *dominance*. Recently, appraisal theories have also been adopted in computational emotion analysis. The approaches differ in their choice of appraisal dimensions: Hofmann et al. (2020) adapt the six dimensions proposed by Smith and Ellsworth (1985) (*pleasantness*, *self responsibility*, *self control*, *effort*, *attention*, *certainty*), while (Troiano et al., 2023) aggregate appraisal dimension form multiple theories into one set of 21 dimensions.

The most prominent task in computational emotion analysis is emotion classification, i.e., assigning emotion representations to textual units. Therefore, emotion analysis is a significantly more complex task than sentiment analysis, as, in NLP, sentiment analysis

is defined as associating text with the basic affect labels of *positive*, *negative* and *neutral*. However, emotion analysis also comprises further tasks such as emotion intensity prediction (Mohammad and Bravo-Marquez, 2017), and emotion role labeling (Bostan et al., 2020 Mohammad et al., 2014). Emotion role labeling is a complex task that identifies the emoter, the target (towards whom the emotion is directed), the cue (lexical indicator of the emotion) and the stimulus of an emotion in the text.

In addition to supporting the analysis of how emotions are conveyed in text, computational emotion analysis opens up a variety use cases. One popular application of emotion analysis is opinion mining, i.e., assessing public opinion towards certain topics or events. In this context, Zhan et al. (2022) investigate which aspects of the COVID-19 pandemic elicit emotional reactions in reddit users. Chen (2013) analyze patients' opinions and emotions towards their medication, doctors and family members as they undergo medical treatment. Further, emotion analysis is applied for assessing argument persuasiveness (Lukin et al., 2017), or for detecting abusive language in online forums (Safi Samghabadi et al., 2020) or hate propaganda in extremist online groups (Abbasi and Chen, 2007). A large number of computational approaches is concerned with developing dialog systems capable of detecting and generating emotional language. The aim of such systems is to improve human-computer interaction through more natural, i.e., emotional, conversation (Song et al., 2019; Rashkin et al., 2019; Herzig et al., 2016). This is particularly relevant for systems that are meant to interact with emotionally or mentally unstable users (Liu et al., 2021).

2.3.1 Methods

Emotion Classification. From the computational perspective, emotion classification has mainly been approached by applying either dictionaries, which rely on predefined word-emotion association pairs, or supervised machine learning, trained on labeled data. In emotion dictionaries, emotion-related terms are associated with a corresponding label indicating their affective information. In WORDNET-AFFECT, Strapparava and Valitutti (2004) enrich entries from WORDNET with various psychological concepts from emotion research, such as bodily symptoms (labeled *emotional response*, e.g., "tremble"), feeling

(*sensation*, e.g., "coldness"), cognitive appraisal (*cognitive state*, "confusion") or basic emotions ("anger"). Esuli and Sebastiani (2006) compile SENTIWORDNET, which scores dictionary entries according to their polarity, i.e., how positive, negative or objective they are. The entries int the NRC lexicon (Mohammad and Turney, 2010) are annotated for Plutchik's eight basic emotions and polarity (positive-negative) in a crowd-sourcing setup.

While dictionary-based approaches can be directly applied to the task, supervised machine learning methods require training on data labeled for the specific task. However, machine learning-based classifiers such as Maximum Entropy (Quan and Ren, 2010), Naive Bayes (Ciobotaru and Dinu, 2021) or Support Vector Machines (Pool and Nissim, 2016) are capable of modeling more complex features than dictionary-approaches. Quan and Ren (2010) explore unigram, n-gram and part-of-speech features for recognizing word emotions in Chinese. Further, they integrate emotion dictionaries into machine learning by considering the word-emotion associations for context words (similarly, Mohammad, 2012). Beyond these generic features, domain- and task specific features are available as well. Mohammad et al. (2015) use hashtags, emoticons and punctuation for classifying emotions from twitter data. Based on psycholinguistic research on emotions, Zanwar et al. (2022) encode morpho-syntactic complexity, lexical richness, diversity and sophistication and readability as features.

Since the rise of deep learning, neural approaches have proven superior to manual feature selection. Among the most common neural architectures for emotion classification are (bidirectional) Long Short-Term Memory Units (Bi-LSTM; Zhou and Wu, 2018) and Convolutional Neural Nets (CNN; Zanwar et al., 2022). The input features in the text are encoded by (contextualized) word embeddings, such as GloVe (Pennington et al., 2014), e.g., in Islam et al. (2019). More recently, transformer-based pre-trained language models, most importantly BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019) and its optimized version RoBERTa (Liu et al., 2019) have been found to consistently outperform previous state-of-the-art approaches in NLP. In BERT, this gain in performance is leveraged by the masked language model pre-training, i.e., the task of recovering masked tokens in the input. This enables BERT to incorporate deep bidirectional context information, which is highly relevant for virtually any NLP task. During fine-tuning, all learned parameters are adapted to the respective downstream task. Since the feature

weights are not trained from scratch, the time and resource requirements for fine-tuning BERT are comparably low.

In the context of emotion classification, Demszky et al. (2020) find that fine-tuning BERT with a comparably simple classification layer outperforms the more sophisticated BiLSTM. Adoma et al. (2020) investigate the performance differences specifically between transformer-based models. Their RoBERTa based approach ranks first, followed by XLNet and BERT.

Appraisal classification. Compared to emotion classification, the task of appraisal classification is still a relatively novel field in emotion analysis. In consequence, the available datasets annotated for appraisal dimensions are sparse, and – analogously to emotion analysis – differ in their choice of appraisal categories. One of the earliest appraisal corpora stems from Hofmann et al. (2020), who re-annotate the existing enISEAR corpus of event descriptions (Troiano et al., 2019) for the appraisal dimensions *attend*, *certainty*, *effort*, *pleasantness*, *responsibility*, *control* (self) and *circumstance* (situational control). The crowd-enVENT (Troiano et al., 2023) corpus combines annotations for both emotions and 21 appraisal dimensions. Both crowd-enVENT and enISEAR comprise annotations for emotions and appraisals, thus enabling investigation of the relation between emotions and appraisals. Hofmann et al. (2020) leverage the twofold annotation for classifying emotions and appraisals in a multi-task learning setup as well as in a pipeline setting, classifying emotion based on predicted appraisal dimensions.

Recently, Stranisci et al. (2022) compiled APPReddit, the only corpus among the here mentioned that does not consist of event descriptions, but reddit posts. It is annotated for five appraisal dimensions, *unexpectedness*, *consistency*, *certainty*, *control* and *responsibility* (all of the mentioned corpora are also considered here. Refer to subsection 3.4 for a more detailed overview.)

From a computational standpoint, there is little difference between the classification of appraisals and emotions. In a multi-task learning setup, Hofmann et al. (2020) classify both emotions and appraisals, where the hidden layers for both tasks are shared; Troiano et al. (2023) base their models on RoBERTa. However, one significant difference between

emotion and appraisal analysis lies the classification setup: Contrary to emotion analysis, where one text is usually annotated with a single emotion (single-label), only the ensemble of all appraisal dimensions combined provides a meaningful representation of the cognitive appraisal. In consequence, appraisal classification is exclusively formulated as a multi-label classification task (or, if investigating to which degree each appraisal dimensions holds, as a multi-label regression task.

2.3.2 Domain Dependency

Independent from the computational approach used, there are multiple challenges for emotion classification, most of which are related to the modality of text: Other than in speech or image recognition, emotion classification from text cannot make use of extra-linguistic features such as prosody or facial expressions. Sarcasm, for example, usually requires visual and prosodic cues to be interpreted correctly [5] (Ray et al., 2022). This also affects to implicit emotions: If no explicit emotion words (or words that bear an emotional connotation) are available, emotion classification has to rely on the context (Ghosal et al., 2021). Alternatively, in texts describing emotion-eliciting events, the underlying emotion can be inferred from the appraisal of the respective event (Klinger et al., 2018). Further, one text might combine the perspective of multiple emoters which requires to either restrict the classification to one single perspective or to account for multi-label classification. The same applies if both the reader's and writer's perspective should be considered (Ramos et al., 2022; Buechel and Hahn, 2017b; Chang et al., 2015).

In order to study challenging phenomena such as the above in isolation, computational emotion classification is thus performed on domain-specific corpora. However, domain-specificity impairs the generalizability of computational models. In order to investigate the differences between in- and cross-corpus classification, Bostan and Klinger (2018) compare 14 popular emotion datasets. The investigated corpora differ in multiple dimensions, most importantly in granularity (i.e., the textual units sharing one annotation; e.g., one sentence or one tweet), emotion annotation scheme (which emotion theory is adopted; which

[5] There are certain lexical cues that authors might use in order to emphasize sarcasm or other non-literal meaning, for example character repetitions to elongate words (Brody and Diakopoulos, 2011).

emotion labels are considered; multi- or single-label), label distribution and domain (e.g., news, fairy tales). In order to enable a systematic comparison despite these considerable conceptual differences, they map the individual annotation schemes onto another, resulting in one overall dataset with unified annotation scheme. In order to enable comparisons between multi- and single-label corpora, they combine the predictions of multiple binary classifiers.

Their results confirm that in a cross-corpus setting, performance drastically decreases compared to the in-corpus classification. This applies especially to those corpora that are from fundamentally different domains (e.g., training on a corpus of self-reported event descriptions and evaluating on a corpus of literary texts) or intended for different tasks (e.g., between a corpus for emotion stimulus detection and another for general emotion classification), Further, they find that some corpora are easier to classify than others (regardless which corpus the classifier was trained on) and that some corpora are more informative than other, in the sense that classifiers perform better in the cross-corpus setting when trained on those. However, Bostan and Klinger (2018) do not conduct further analyses to investigate what (besides the corpus-specific differences mentioned above) is causing the differences in performance.

Since there are considerably less appraisal than emotion corpora, no such large-scale analysis of cross-corpus or cross-domain classification has yet been conducted for appraisals. However, Stranisci et al. (2022) provide a similar comparison on a smaller scale by mapping the annotation scheme of their APPREDDIT corpus to ENISEAR. Despite the conceptual differences between the two corpora, Stranisci et al. find that they can indeed be combined without significantly decreasing the classifier's performance. However, they align their aggregated corpus only along four dimensions while the remaining four (*unexpectedness* from APPREDDIT, *attend*, *effort*, *self responsibility* and *other responsibility* from ENISEAR) could not be aligned and are therefore not considered.

2.3.3 Domain Adaptation

Emotion classifiers thus cannot be easily adapted from another task or domain. Instead, they need to be trained and tuned on each specific dataset individually. If no such training

data is available, it might be necessary to compile and annotate a novel dataset, a task that is very costly in terms of resources and time [6].

In order to avoid such efforts, transfer learning methods aim to adapt a model from one domain to another, unlabeled domain. This is generally achieved by learning basic features relevant to the respective task, while ignoring other, domain-specific features, which results in a domain-agnostic classifier. One approach to obtaining these features is through pivot features, i.e., terms that frequently occur in data from both domains and share the same emotional connotation across domains (e.g., *loved it*; cf. Blitzer et al., 2007). Unlabeled texts from the target domain can then be classified based on their correlation with pivot features, based on the assumption that terms occuring in the context of the same pivot features share the same emotional connotation. An alternative way to learn domain-agnostic features is through adversarial learning. Adversarial methods are not exclusive to domain adaptation, but can be applied to a variety of tasks and domains (cf. Wang et al., 2019). For domain adaption, Ganin et al. (2015) proposes a gradient reversal architecture (3): A standard, task-specific classifier is complemented by an adversarial domain-discriminator, tasked with predicting the original domain of the instance (Li et al., 2019; Du et al., 2020). During the forward pass, the gradient reversal layer has no influence on the classifier, but during backpropagation, it inverts the gradient (and multiplies it by a hyperparameter λ). Therefore, domain specific features supporting the task of domain classification are suppressed, but task-specific features are encouraged. Since the adversarial task of predicting domain constrains the domain-specific features, no manual selection of pivot features is required.

What has not been considered yet, however, is whether the domain-specific features that hinder cross-domain emotion classification consists – at least in part – of topic-related features. Here, I hypothesize that a major part of what makes a domain unique is the distribution of topics within, besides style-related features, such as syntactical features: The emotions associated with the topic of *guns* arguably diverge between corpora from different domains, e.g., between a corpus of gun violence reports (such as Reardon et al., 2022) and a corpus of social media posts. Under this assumption, it needs to investigated

[6] Although the quantity of data required has decreased with the introduction of pre-trained models (cf. 2.3.1), the need for task- and domain-specific data still persists (cf. Du et al., 2020).

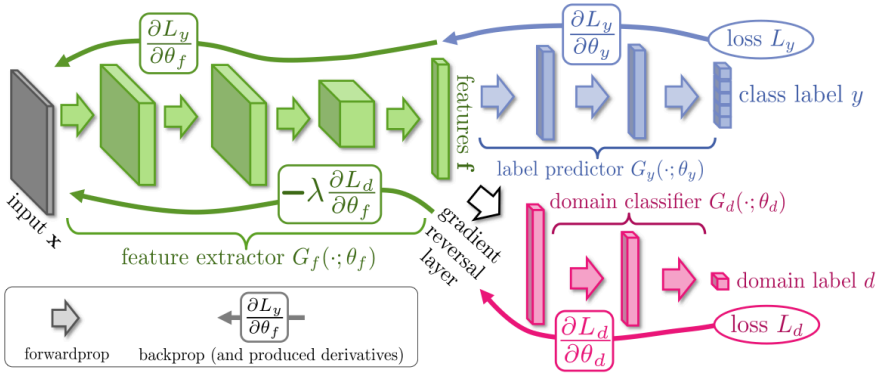

Figure 3: Gradient Reversal Layer for domain adaptation as proposed by Ganin et al. (2015).

to what degree reducing the influence of topic-related features in emotion classification contributes towards improved cross-domain performance.

2.4 Bias

In the context of NLP, bias has been found to affect computational resources of various kinds. As such, it is a popular research subject not only in computational analysis but also in hate-speech-detection (Wich et al., 2020), sentiment analysis (Wang et al., 2021), machine translation (Stanovsky et al., 2019) or argument mining (Spliethöver and Wachsmuth, 2020a). In general, the term "bias" refers to the phenomenon that machine learning models adopt latent, "non-generalizable features" (Shah et al., 2020) from the training data, such as domain-specific terms, contexts or text styles. If the training data lacks demographic variation, the model then adopts the learned representation as the standard, i.e., it develops a bias. In consequence, the biased representation leads to erroneous results when applied to a domain where the alleged standard does not hold (cf. Hovy and Prabhumoye, 2021).

Bias is typically arises from demographic features in the text, such as age, gender or ethnicity, which are conveyed beyond the literal text meaning. For example, the two formulations "I am totally pumped" and "I am very excited" share the same basic meaning, but additionally imply information about the authors' age, and social status (adapted from Hovy and Prabhumoye, 2021). The issue of bias is thus inherent to language, and in consequence, to NLP: Bias is present in word embeddings (Basta et al., 2019), language models (Nadeem et al., 2021), as well as in task-specific models and resources (e.g., for hate-speech-detection, Davidson et al., 2019). However, the exact definition of the term "bias" is ambiguous and differs between approaches (Blodgett et al., 2020). In general, two different perspectives on bias can be distinguished.

Impacts of bias. One way to approach the issue is by assessing the harmful potential of biased models. If a model is trained on data which contains stereotypical representations, these stereotypes are adopted by the model. Such social bias is directed towards certain demographic groups, defined by gender (Sun et al., 2019), ethnicity (Davidson et al., 2019), age (Fraser et al., 2022) or physical features, such as disabilities (Hutchinson et al., 2020). One well-known issue in this context is the incorporation of gender bias in word embeddings; illustrated e.g., by Bolukbasi et al. (2016) who find the relation *man – woman* to be equivalent to *computer programmer – homemaker* in word embeddings trained on news data. The issue is also compounded by the fact that word embeddings further propagate the bias through their widespread application in other NLP models. Social bias is also present in emotion and sentiment analysis, e.g., when male and female emoters are assigned different emotions or different emotion intensities although the context is otherwise identical (Kiritchenko and Mohammad, 2018). Diaz et al. (2018) investigate the impact of age bias in sentiment analysis.

In order to categorize bias and the harm it causes, Crawford distinguishes between *allocation* and *representation* bias (cf. Sun et al., 2019) In allocation bias, systems exhibit a bias towards minority groups by performing better on data associated with the majority class. For example, speakers of minority languages are affected by allocation bias in the sense that machine translation systems are typically underperforming when translating to or from a minority language. Representation bias denotes the skewed associations contained

within computational representations of text, as depicted in the examples of gender bias above. Dev et al. (2022) approach the issue by introducing a framework for assessing the harms caused by a biased system. The measures comprise of stereotyping, disparagement, dehumanization, erasure (lacking representation of certain groups) and quality of service (equivalent to Crawford's allocation).

Sources of bias. Another perspective on bias focuses on how bias is introduced to datasets and models in NLP Dey et al., 2020. While still taking the impact of bias into account, Hovy and Prabhumoye (2021) primarily investigate the "mismatch of ideal and actual distributions of labels and user attributes in training and application of a system" (p. 2). They identify five steps in the development of NLP applications that bear the potential of introduction bias to the system:

The **selection of training data** is, as illustrated above, one of the most impactful decisions concerning bias. Due to the domain-dependence in NLP (cf. 2.3.2), training data often lacks the diversity required in order to account for different demographics. Many widely used resources involved in the development of NLP systems (i.e., datasets, word embeddings, syntactic parsers, etc.) are based on linguistically and demographically homogeneous – and partially outdated – domains, e.g., traditional newspapers from the 1990s (cf. Hovy and Prabhumoye, 2021; p. 4).

Even if the data is carefully chosen, bias can still be introduced through the **annotation procedure**. The decisions involved in annotating corpora depend on the individual annotator's personality and beliefs, especially if the annotation task is inherently subjective[7]. A homogeneous group of annotators or unqualified non-expert annotation might thus lead to bias in the labeled dataset (Sap et al., 2022; Biester et al., 2022). Beyond data aggregation, the **input representation** as well as the **computational model** are primary sources of bias, specifically the bias encoded in word embeddings and the tendency of machine learning algorithms to pick up on semantically unrelated, but statistically related features.

[7]Due to the subjective nature of emotions, this is especially relevant for annotating emotion datasets. In this context, Milkowski et al. (2021) assess annotators' individual *Personal Emotion Bias*. They find that informing an emotion classifier with this personality-specific information improves the classification of controversial instances (i.e., instances with low inter-annotator agreement).

Further, Hovy and Prabhumoye (2021) find the overall **research design** to be a potential source of bias, e.g., when focusing only on a single language (English) instead of considering minority languages.

2.4.1 Topic Bias

As illustrated, the notion of bias in NLP is not consistently defined and thus approached from different perspectives. However, the two perspectives introduced above are not necessarily mutually exclusive and often overlap. This is also reflected in the terminology used around bias: The terms used in the context of social bias, i.e., gender, age, ethnic bias, etc., refer to the demographic group that is impacted by the respective type of bias. In contrast, terms such as authorship bias (Wiegand et al., 2019)[8], annotator bias (Sap et al., 2022), sampling bias (Razo and Kübler, 2020) and, specifically, topic bias, refer to the cause of respective bias. Therefore, terms from both categories can, in principle, be used simultaneously to refer to the same phenomenon: In the case that the bias originates in skewed topic representations, which, in turn, manifest in harmful behavior, both topic bias and social bias are present. In the context of investigating bias in abusive language datasets, Wiegand et al. (2019) find the topic of "soccer" to be almost exclusively associated with abusive language. That topic bias is caused by the sampling procedure applied for that specific dataset, as it features a large number of online posts discussing the role of women in sports – which, as Wiegand et al. (2019) point out, are likely to exhibit abusive, i.e., sexist language. The same phenomenon can thus be referred to by four distinct types of bias: As sampling bias (with respect to the data aggregation that caused the overrepresentation of abusive-labeled "soccer"-instances), topic bias (since the topic of "soccer" is highly correlated with abusive language), overamplification bias (as defined by Hovy and Prabhumoye, 2021; meaning that the bias in the data is picked up and amplified by a model trained on it) and gender bias (referring to the sexist stereotypes adopted from the source domain).

[8] Authorship bias might emerge if a significant amount of instances is authored by a single individuum. Even if author information is not explicitly encoded through additional features, a classifier might still learn to associate the individual writing style or preferred topics with the respective class, instead of considering the actual features in the text (cf. Wiegand et al., 2019).

In this work, topic bias is understood to comprise two of these concepts: First, the association of certain emotion or appraisal labels with certain topics and second, the resulting bias in a classifier towards certain topics when predicting the emotion and appraisal labels. As the investigation conducted here is rather foundational, the impact, i.e., potential harms caused by topic bias in emotion classification are not considered in further detail.

2.4.2 Bias Detection

Since bias can be introduced at various points in the development of an NLP system, the methods proposed for detecting bias are also specific to the respective components within such system.

For detecting bias contained within **pre-trained models and word embeddings**, Caliskan et al. (2017) introduce the Word Embedding Association Test (WEAT), which is based on the mean cosine similarity between certain target words (e.g., European American and African American names) and corresponding association words (e.g., pleasant and unpleasant). Kurita et al. (2019) introduce a method for assessing biased word embeddings in BERT: Based on the masked language model task applied during pre-training, they generate sentence templates, consisting of a gender-specific masked target term (e.g., gendered pronoun) and a stereotypical attribute (e.g., "[MASK] is a programmer"; cf. **?**; p.167). Comparing the probabilities that BERT assigns to either the male ("*he* is a programmer") or female ("*she* is a programmer") instantiation of the masked target reveals the stereotypical biases within the masked language model. Similarly, Huang et al. (2020) prompt a language model to generate various continuations for a given conditioning sentence. They find that certain attributes in the conditioning sentence, such as stereotypical male or female professions, influence the sentiment of the generated continuation, thus displaying bias.

This work here is specifically related to approaches detecting bias in **datasets** and **task-specific classifiers**. In order to assess bias in a corpus of abusive language, Wiegand et al. calculate the pointwise mutual information between words and the *abusive* annotation. They find that most of the strongest correlated words are not by themselves indicative of

abusive language (e.g., “commentator” or “football”). Instead, the high correlation arises from the biased distribution of *abusive* labels towards these topic words. They investigate further datasets and find similar correlations, e.g, indicating racist bias (high correlation with Arabic names). In order to assess the impact on modeling, they train a classifier on the biased data and evaluate it on in a testset which contains some of the biased topic words, but only very few actually labeled as abusive. Wiegand et al. find that the classifier reflects the bias in the training data by misclassifying a high amount of texts as abusive. Their findings are especially relevant since the biased dataset yields best in-corpus classification results among all of the datasets considered in their analysis. They infer that the good classification performance, especially on implicit abusive language, is due to the classifier learning features associated to the correlated topics instead of features actually relevant to abusive language task. To further assess how a non-biased classifier would perform, Wiegand et al. remove a portion of the highly correlated (but unrelated to the task of abusive language detection) terms from the dataset. They report that the classification performance consequently decreases.

Similarly, Nejadgholi and Kiritchenko (2020) also investigate topic bias in abusive language datasets and how its influence on cross-corpus classification. However, they differ in their methodology from Wiegand et al. (2019): Instead of calculating pointwise mutual information, they train a topic model on one of the datasets and perform a qualitative analysis of the resulting topics. In order to assess the differences in topic distribution across different datasets, Nejadgholi and Kiritchenko apply the topic model to subsets of other abusive language datasets and evaluate a classifier (trained on the same dataset as the topic model) on each of the topics defined by the topic model. Their results show that, while the classifier’s performance is quite high overall, performance varies significantly between topics. This behavior again confirms that class labels (at least in the case of abusive language) are in fact not equally distributed across topics and thus biased.

Devinney et al. (2020) also apply topic modeling, but for investigating gender bias across multiple datasets. Other than Nejadgholi and Kiritchenko, they use a semi-supervised topic modeling approach, which allows them so define certain seeding terms that foster clustering of related terms. Since they are investigating gender bias, they use gender-specific terms for seeding (e.g., “woman” or “man”), in order to prompt the model towards

gender-centric topics. They assess gender bias by analyzing the term distribution across topics and compare it between datasets. This quantitative analysis allows to infer how the different genders are represented in each corpus. For example, Devinney et al. find that all datasets exhibit outdated gender roles, as the feminine-gender class is associated with topics such as “family” and “home”. However, their approach is limited to a purely qualitative analysis and they do not provide a modeling approache to test whether the bias is also reflected in classifiers trained on the data.

Other approaches aim to reveal the bias incorporated in computational models through specific testsets. Kiritchenko and Mohammad (2018) compile the Equity Evaluation Corpus (EEC), which consists of sentence pairs differing in only one word bearing either gender or ethnic information. On this corpus, they evaluate 219 different models for emotion intensity prediction in order to test whether the predictions differ between the sentence pairs. Barikeri et al. (2021) construct RedditBias, consisting of reddit comments annotated for various dimensions of social bias. Based on this dataset, they propose a method for evaluating bias in language models by measuring the perplexity between a biased sentence and a corresponding, inversely biased sentence.

2.4.3 Bias Mitigation

Analogous to bias detection methods, approaches for bias mitigation tackle the issue at different stages of the NLP system. Mostly, it can be differentiated between methods that address bias at the data level (i.e., correcting the initial cause) or at the modeling level (correcting the impact).

Mitigating bias in data. For the former, Wiegand et al. (2019) propose manual debiasing. In the case of topic biased data, this involves sampling additional texts of the same topic at random, while the number of texts per author could be restricted in order to prevent author bias. This is motivated by their investigation of abusive language datasets, revealing that biased sampling strategies are the main cause for introducing bias. For example, bias is introduced through the query terms that were used to sample the data: Since Wiegand et al. find some of them highly correlated with abusive language while others are not,

the topic-specific query terms (e.g., “WomenAgainstFeminism”) lead to a skewed label distribution across topics.

Another debiasing method that is directed towards the biased data itself is proposed by Barikeri et al. (2021). In order to account for an overrepresentation of racist bias, they augment the training data with counterfactual instances. Each biased sentence is duplicated, replacing the biased term with the inverse term in the duplicate (e.g., “that Muslim is dangerous” is augmented by “that Christian is dangerous”). This procedure thus prevents the model from picking up bias towards a specific social (minority) group.

Mitigating bias in models. While manual debiasing allows for training potentially bias-free models, it is both time and resource intense which might render it infeasible especially for large datasets. Instead, debiasing could also take place during training, rendering manual manipulation of the training data obsolete. This can be accomplished through adversarial learning, which prevents a neural model from learning certain information that might introduce bias to the model. Originally introduced for domain adaptation (2.3.3), the gradient reversal architecture has also been applied to other fields of research, including Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Medical Records in **?**. For mitigating bias, the structure is adapted so that the adversarial classifier (cf. 3) ,originally tasked with discriminating between domain, predicts the topic of the respective instance. The remainder of the architecture remains unchanged, so that learning of bias (e.g., topic) features is discouraged.

An alternative setup is proposed by **?**, who first train an intentionally biased classifier in order to identify the features that exhibit bias. This information is then used to train the debiased classifier which compensates for the features learned by the biased classifier (DRiFt). Further, **?** adapt the language model’s loss function in order to mitigate gender bias. They introduce a new term to the loss function that aims at equalizing the probability of male and female words.

The related work on bias as well as on the detection and mitigation of bias that has been presented in this section illustrates two key aspects:

First, there is strong evidence from multiple research areas that datasets are, in fact,

biased. While a considerable amount of the presented research tackles gender and social bias (due to the general predominance of research on these types of biases), the works that specifically address topic bias reach the same conclusion.

Second, bias has yet been addressed in the context of emotion analysis only for social bias[9]. This reveals a research gap concerning the relation of topics and emotions, which this work is addressing.

3 Methods

The following section provides the formal definition of the here presented investigation on topic bias and details the experiments that are conducted to address the initially formulated research questions. Further, 3.2 details the topic modeling method applied to obtain the topic labels, while 3.3 summarizes further implementation details. The investigated corpora are presented in 3.4.

3.1 Experimental Setting

This investigation considers multiple corpora, where each corpus $c \in C$ is modeled as a tuple consisting of a set of topic labels T_c, a set of instances I_c and a set of annotation labels L_c, where L_c is either sampled from the set of overall appraisals ($L_c \subseteq A_C$) or emotion labels ($L_c \subseteq E_C$), where $A_C \cap E_C = \varnothing$[10] .

Further, each instance $i_c \in I_c$ consists of a text $s_{i,c} = (s_1, s_2, \ldots s_n)$, a topic label $t_{i,c} \in T_c$ and a set of emotion or appraisal labels $L_{i,c} = \{a_j, \ldots, a_k\} \subset L_c$. Since some of the corpora investigated here are labeled with multiple, i.e., one or more emotions, the number of emotion labels is defined by $1 \leq i, j \leq |L_{i,c}|$. Appraisals are always annotated

[9] Based on the differing definitions of the term "bias" (2.4.1), it might be possible that topic bias has been implicitly part of previous work on other bias types (e.g., gender) in emotion analysis . However, the here presented approach is the first to address topic bias explicitly and in full detail.

[10] Since CROWD-ENVENT is annotated for both appraisals and emotions, this formal definition considers it as two corpora with identical topic and text sets, but differing labels.

C	ISEAR	$s_{\text{i,ISEAR}}$	*When as a 7 year old, I stole money from my mother.*
L_{ISEAR}	{*anger*, *disgust*, *fear*, *joy*, *sadness*, *shame*}	$L_{\text{i,ISEAR}}$	{*shame*}
T_{ISEAR}	{*love, exams, death, shame, school, animals, alcohol, accidents, fear, theft*}	$t_{\text{i,ISEAR}}$	*theft*

Table 1: Example instance from the ISEAR corpus. As ISEAR is a single-label emotion corpus, the set $L_{i,c}$ only contains one label.

in a multi-label setting, which is why $6 \leq i, j \leq |L_{i,c}|$ (since six is the minimum number of labels that any of the appraisal corpora is annotated for). An example instance i is provided in Table 1.

While the emotion (E_C) and appraisal annotations (A_C) are already included in the respective corpora, the topic labels T_C are generated specifically for this investigation. They are obtained by training topic models on each individual corpus (refer to 3.2 for a detailed description of the applied topic model).

3.1.1 Analysis of Bias in Emotion Datasets

In order to address the first research question, whether datasets in emotion analysis exhibit topic bias, I build on top of previous work on analyzing topic bias in datasets. Precisely, I follow the topic modeling approach proposed by, i.a., Nejadgholi and Kiritchenko (2020) and Devinney et al. (2020), who conduct qualitative analyses of topic distributions. Here, the analysis is divided into two perspectives, namely at the corpus- and the instance-level:

Analyzing topic distribution First, I analyze the set of corpus-specific topics T_c, i.e., the general distribution of topics both within and across corpora. This analysis provides first insights on how topics reflect the individual domain of each corpus.

Analyzing topics and emotions Second, I considering the instance-specific topic labels $t_{i,c}$. By investigating the correlations between topics and emotion (or appraisal) annotations, I provide a qualitative analysis of topic bias in emotion datasets.

I hypothesize that emotion corpora are as diverse in topic as they are in domain, and that certain topics are prototypically associated with certain emotions. These analyses serves as the basis for the subsequent experiments, as they provide a first assessment of topic bias in the investigated corpora.

3.1.2 Detecting Bias in Emotion Classifiers

For investigating the impact of biased datasets on the resulting classifiers, I mainly follow the methods applied by Wiegand et al. (2019) and Nejadgholi and Kiritchenko (2020) (cf. 2.4.2), who train and evaluate classifiers across subsets sampled from different topics. The following experiment is thus targeted at revealing potential topic bias within the discussed datasets and the resulting classifiers.

For each topic $t_c^{out} \in T_c$ in a given corpus, I train separate classifiers tasked with predicting either the emotion or appraisal label $a \in L_{i,c}$. In the subset corpus used for training the classifier (T^{train}), instances with the topic label t_c^{out} are excluded, i.e., $T_c^{train} = \{t_{i,c} | t_{i,c} \in T_c, t_{i,c} \neq t_c^{out}\}$. The number of classifiers trained for a given corpus c is thus equal to $|T_c|$. I refer to the topic t^{out} as the *held-out* topic of the respective classifier. The classifiers are evaluated in two distinct settings:

INTOPIC In the first setting, multiple testsets are sampled from the corpus, one for each topic except t_c^{out}. Each testset is thus defined in relation to the respective held-out topic: $t_c^{in} = T_c \setminus \{t_c^{out}\}$. Thus, the union of all t_c^{in} per corpus reflects T_c^{train}. Therefore, a classifier trained on T_c^{train} is evaluated on all t_c^{in} of corpus c.

CROSSTOPIC For the second setting, the classifier is evaluated on the held-out topic t_c^{out} which is not part of the training set T_c^{train}.

Comparing the differences in performance between the INTOPIC and CROSSTOPIC setting, allows to assess whether the classifier is biased towards the topics seen during

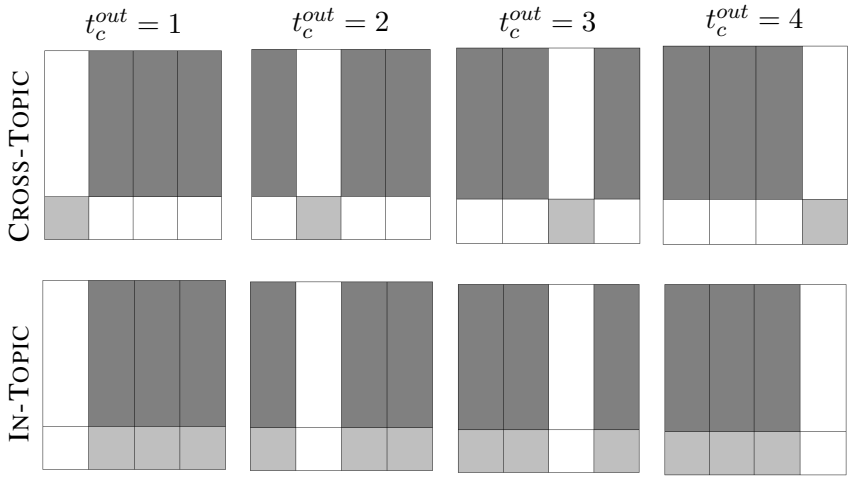

Figure 4: Exemplary setup of INTOPIC and CROSSTOPIC experiments on a corpus c with four topics. For each experiment, ■= train split, ■= test split.

training: If the classifier is not influenced by topic, I expect the results of INTOPIC and CROSSTOPIC to be on par. In that case, the t_c^{out} testset in CROSSTOPIC would no different than a regular held-out subset used for standard evaluation. However, a decrease in performance in the CROSSTOPIC setting (as compared to INTOPIC) would indicate that the classifier relies on topics for the emotion classification and misclassifies instances if met with unseen topics during testing. Figure 4 provides an overview over both settings[11].

3.1.3 Mitigating Bias

To address the research question of whether emotion classifier can be de-biased with respecto to topics (provided that evidence for the existence of topic bias is found), two experiments are conducted.

Topic word removal The first experiment is a naive approach to topic bias mitigation based on Wiegand et al. (2019): In order to reduce the influence of topic words on

[11]Note that the illustration in 4 is simplified. In the actual implementation of all in- and cross-topic experiments, each subset t_c^{in} as well as the corresponding t_c^{out} are split into a train-, test- and validation-subsets to prevent the classifier from overfitting on the in-topics during training.

classification, the respective words are removed from both the training- and testsets. However, the experiment carried out here is not identical: While Wiegand et al. remove words that they find to be most correlated (but not semantically related) with a certain annotation label, here, the most indicative words for each topic are removed (based on the probability the topic model assigns to each word). This difference is due to the slightly different research questions: Wiegand et al. aim to assess the performance of a classifier trained on the de-biased dataset, while the goal of the present experiment is to assess whether de-biasing in emotion classification is possible at all. The de-biased classifier is then trained and evaluated in the same setting as described above, i.e., in the CROSSTOPIC and INTOPIC setting. I thus refer to this experiment as INTOPIC-MASK and CROSSTOPIC-MASK, respectively.

Gradient Reversal The second experiment represents a more sophisticated approach to de-biasing. Based on common approaches to bias mitigation, I extend the emotion/appraisal classifier by a topic predictor and gradient reversal layer. Again, the approach is tested in the INTOPIC and CROSSTOPIC-MASK setting (INTOPIC-GRL and CROSSTOPIC-GRL).

Compared to the biased classifiers in the experiment on bias detection, I hypothesize that the de-biasing methods lead to a decrease in performance overall, since less features will be available to base classification decisions on. However, at the same time, I expect the difference in performance between INTOPIC and CROSSTOPIC setting to decrease as well, since the reason for why the INTOPIC setting is supposedly more performant than the CROSSTOPIC setting is mitigated (i.e., the topic bias caused by the topic-specific words).

3.2 Topic Modeling

The topic modeling method has the most far-reaching consequences for the entire investigation, which is why the process is presented here in more detail.

The most established topic model is Latent Dirichlet Allocation (LDA; Blei et al., 2003), which has also been applied for investigating topic bias (Nejadgholi and Kiritchenko, 2020). Since LDA is a Bayesian model, it models documents as bag-of-words representations. In

consequence, it cannot account for context, i.e., it does not model the relation of words within sentences.

Therefore, I consider an alternative approach to topic modeling in addition to LDA, namely BERTopic (Grootendorst, 2022), which leverages pre-trained transformer models to model the semantic relations within sentences and has been proven effective in previous research (Xu et al., 2022, Kellert and Mahmud Uz Zaman, 2022, Eklund and Forsman, 2022). A further advantage of BERTopic over LDA is, that BERTopic offers HDBSCAN as a clustering method, which does not require a pre-determined number of topics. Instead, clustering is based on document similarity, ignoring outliers.

I conduct a systematic comparison of both techniques in order to achieve a meaningful topic model that the subsequent experiments can be build upon.

3.2.1 Implementation

LDA requires the number of topics as a hyperparameter. I start with the number of topics that were identified in the manually annotated subset and continue hyperparameter search from there. I also consider the same number of topics identified by BERTopic. I report the configuration that yields best results in the evaluation below. Apart from the choice of number of topics, LDA is implemented as in the sklearn library (Pedregosa et al., 2011) with default parameters left unchanged. For the features extraction, tf-idf-features are used, general stop words as well as corpus specific stop words (words appearing in over 90% of all documents) are removed.

BERTopic consists of a pipeline of components for features representation, dimensionality reduction, clustering and topic extraction, each of which can be modified. Here, I use a pre-trained sentence-emdedding (all-MiniLM-L6-v2, as implemented in Huggingface) for feature extraction, Accelerated Hierarchical Density Clustering (HDBSCAN; McInnes and Healy, 2017), Uniform Manifold Approximation (UMAP; McInnes et al. (2020)) for dimensionality reduction and tf-idf for retrieving the topics within the clusters. Although HDBSCAN does not require a pre-determind number of topics, it can be tuned by setting hyperparameters for the minimum cluster size and controling the amount of outliers allowed within a cluster. I adapt these hyperparameters to each corpus individually,

depending on its size. Again, the results reported in the evaluation are with hyperparameters that yielded best results.

3.2.2 Evaluation

In order to enable an extrinsic evaluation of approaches, a set of gold topic labels is prepared for each corpus. This gold annotation is compiled by manually annotating 50 instances randomly sampled from each corpus (by the author or this work). Since the limited number of only 50 instances per corpus cannot account for all possible topics within a corpus, the evaluation can only serve as an approximation of the models' actual modeling capabilities. However, since a larger scales annotation study would have been out of the scope of this work, this evaluation method is an acceptable compromise.

The evaluation includes multiple metrics proposed for assessing the quality of clusters with respect to a gold standard: Rand index, adjusted Rand index (adjusted for agreement by chance), homogeneity score, and V-measure (combination of homogeneity and completeness of topic clusters). Further, I also report F_1.

3.2.3 Results

The results for both models on each corpus, as well as the delta for each metric (BERTopic to LDA) are reported in Table 2. Notably, across all corpora, the BERTopic approach clearly outperforms the LDA method for F_1 and adjusted Rand index. The Rand index, however, is more similar between both methods, and LDA outperforms BERTopic in four out of the six corpora. I hypothesize that this is due to the strong impact of bigger clusters on the calculation of Rand index. If considering the adjusted Rand index, which accounts for random agreement between predicted and gold-standard clusters, LDA is considerably and consistently outperformed by BERTopic. The Tales corpus appears to be the most challenging to cluster for BERTopic. However, this is also due to how the gold labels were annotated: Due to the literary domain of the sentences in Tales, it was often challenging to determine topics among the 50 investigated instances (e.g., "One bell was silent; but it was illuminated by the bright sunshine which streamed from the head and bust of the renowned figure, of which it formed a part.").

	LDA					BERTopic					Δ				
Corpus	V	H	RI	ARI	F_1	V	H	RI	ARI	F_1	V	H	RI	ARI	F_1
ISEAR	49	50	83	7	16	61	59	84	22	31	+12	+9	+1	+15	+15
SSEC	37	42	76	2	13	73	78	86	46	54	+36	+36	+10	+44	+41
Tales	42	44	80	−2	9	41	33	60	11	30	−1	−11	−20	+13	+21
crowd-enVENT	39	44	76	0	11	43	44	74	8	23	+4	0	−2	+8	+12
APPReddit	47	51	82	0	7	57	52	76	22	36	+10	+1	−6	+22	+29
enISEAR	61	61	90	−1	4	69	63	86	23	29	+8	+2	−4	+24	+25

Table 2: V-measure (V), homogeneity (H), Rand index (RI), adjusted Rand index (ARI) and F_1 for both topic modeling models LDA and BERTopic.

Since the small set of manually labeled instances only provides an approximation to the models' performance, I conduct an additional, intrinsic evaluation by analyzing the most representative words for each topic generated by LDA and BERTopic. As an example, the most representative words (out of space constraints, only the top four are reported here) for ISEAR determined by LDA and BERTopic, respectively, are reported in Table 3 (for all other corpora, refer to Appendix B). Note that the order of topics and associated ID (which is only included for ease of referencing individual topics here) in the table is arbitrary, thus, there is no relation between LDA and BERTopic generated topics of the same ID. The manually determined topic label provided in the second column only applies to the topics generated by BERTopic, since for most of the topics generated by LDA, the semantic variation between the topic words cannot be captured with a single label. In both models, topic ID 0 denotes words that do not seem related and do not fit with any of the other topics. Here, this topic is referred to as the *outlier* topic. In BERTopic, this cluster is purposely generated to ensure a better coherence among the other, "true" topics (as outliers can be removed and clustered in the outlier cluster instead).

The topics generated by BERTopic (based on the most representative terms) are considerably more coherent and, intuitively, more meaningful, which is reflected by the distinct label than can be associated with each set of topics. The topics in LDA, however, appear to contain more noise in the form of outliers. For example, the word "mother" is highly representative for the LDA topics 4 and 6, while semantically related terms ("father",

	BERTopic		LDA
ID	Label	Top 4	Top 4
0	*outlier*	response, discussion, turn, received	told, didn, asked, went
1	love	love, lover, relationship, loved	girl, felt, ill, time
2	exams	exam, exams, examination, examinations	grandfather, old, grandmother, father
3	death	death, funeral, dying, died	man, english, went, like
4	shame	ashamed, guilt, shame, depressed	met, passed, away, mother
5	school	lecture, teacher, lecturer, classroom	home, later, heard, friend
6	animals	dog, dogs, animal, pet	mother, old, years, house
7	alcohol	drunken, drunk, alcoholic, drank	studying, time, got, failed
8	accidents	accident, driving, bus, drove	people, young, got, home
9	fear	frightened, fear, terrified, fearful	disgusted, did, work, examination
10	theft	theft, stealing, stole, thief	went, brother, time, selected
11			people, husband, doing, girl

Table 3: Each topics (ID) most informative keywords (Top 4) found in ISEAR by the BERTopic and LDA topic models with ID 0 denoting the BERTopic outlier class. The manual label for each topic (Label) only applies to BERTopic because of the erratic nature of the LDA topics.

"grandmother", etc.) are clustered in topic 2. Similarly, the semantically related topic words "studying" and "examination" appear in different topics. In BERTopic, on the other hand, the topic of "examination" (ID 2) is very homogenous, as the four most representative words are all highly similar (in fact, they are lexical variations of the same word).

Based consistently better performance of BERTopic over LDA (most importantly, F_1 and adjusted Rand index) as well as the considerably more homogeneous distribution of representative words per topic, only the topics generated by BERTopic are considered in the subsequent analyses. However, the downside to the highly homogeneous topics in BERTopic is, that the reduced number of outliers per topic results in a larger overall outlier topic (ID 0). This is addressed in the context of the qualitative analysis of topics in 4.1.

3.3 Implementation Details

Emotion and Appraisal Classifier Following state-of-the-art approaches to emotion and appraisal classification (Demszky et al., 2020 Troiano et al., 2019), I fine-tune RoBERTa (Liu et al., 2019) as implemented in the Huggingface library (Wolf et al., 2020) on each corpus. For the classification, the output from the transformer layers is pooled and passed through a fully-connected dense layer (768 units). I apply ReLU activation (Agarap, 2019) and a dropout of 0.5. I apply a consecutive classification layer using softmax activation and binary cross-entropy loss for single-class classification (for ISEAR, TALES, and emotions in CROWD-ENVENT). For the multi-class classification task (SSEC, APPREDDIT, ENISEAR and appraisals in CROWD-ENVENT), I apply a sigmoid activation and categorical cross-entropy loss instead. The learning rate is set to 5×10^{-5} across all experiments; the batch size is 16. I train each classifier for a maximum of 5 epochs but apply early stopping based on the validation accuracy (stops after two consecutive epochs without improvement). As optimizer, AdamW (Loshchilov and Hutter, 2019) is applied, weight decay is set to 10^{-5}. Results are averaged over three different runs for each classification task.

Removal of Topic Words The list of topic words to be removed in each corpus is consists of the then most representative words of each topic within the dataset. The most representative words, i.e., the top k words per topic are determined by the probability that BERTOPIC assigns to each word, i.e., the word's probability to be assigned a certain topic label. Therefore, k is a hyperparameter determining the trade-off between general classification performance and topic-influence: Increasing k increases the potential impact of the de-biasing method (as less topic-specific features are available to the classifier), but, at the same time, decreases the general classification as less and less features are available overall. Further, by choosing a higher k, more words which are less representative for a given topic are removed as well, thus introducing noise to the experiment. Here, k is here set to 10 (setting$k = 3$ or 5 was considered as well). This choice is motivated by the observation that the top k representative words often comprise variations of the same word or concept. For example, in ISEAR, the ten most representative words for the topic *theft* consist of "theft", "stealing", "stole", "thief", "robbery", "thieves", "stolen", "borrowed", "robbers" and "cash". A higher k thus covers a broader range of morphological ("stealing",

"stole", "stolen" and "thief", "thieves"), as well as semantic ("theft", "robbery") variation.

Notably, the chosen topic words were not removed from the input. Instead, based on the masking of explicit emotion words in Troiano et al. (2019), they are substituted with "..." before being fed into the RoBERTa tokenizer. The number of masked topic words per corpus is summarized in Table 15 (Appendix A).

Gradient Reversal Layer The gradient reversal layer (GRL) is implemented as described in Ganin et al. (2015). The purpose of the GRL is to reverse the gradient (by multiplying it with $-\lambda$) of the following layer during backpropagation. Since the layer has no trainable (nor non-trainable) weights associated with it, the GRL has no effect during a forward pass and acts as an identify transform. For the InTopic-GRL and CrossTopic-GRL experiments conducted here, the GRL is added into the standard classifier architecture described above. As depicted in Figure 5, the emotion classifier is coupled with an additional topic classification layer, equivalent to the single-class emotion classification layer, with the task of predicting the correct topic label $t_{i,c}$ for each instance. The topic classifier is connected via the GRL to the remaining layers of the network, i.e., the pre-trained RoBERTa model as well as the single dense layer. Since the gradient is reversed, all weights in the shared layer associated with the topic prediction task are decreased. A key factor in the implementation is the choice of λ as it regulates the impact of the GRL. Again, choosing λ is a trade-off between overall classification performance and de-biasing potency. To determine an optimal value for $lambda$, standard emotion (or appraisal classifiers) are trained on each individual corpus for $lambda$ values of 0.1, 0.3, 0.5, 1 and 3. Across corpora, a significant decrease in performance (often scoring 0 F_1 for multiple classes) can be observed for any $\lambda > 0.1$. Therefore, λ is set to 0.1 for all InTopic-GRL and CrossTopic-GRL experiments.

Cross-Domain Classification The classifiers applied in the CrossDomain-GRL experiments are structurally identical to the one classifiers described above. However, CrossDomain-GRL must take different annotation schemes between training and testdata into account, i.e., training on a corpus annotated with single labels while testing on a multi-label corpus (and vice-versa). Here, this is addressed by adapting the method proposed by Bostan and Klinger (2018): When training on multiple labels and testing on single labels

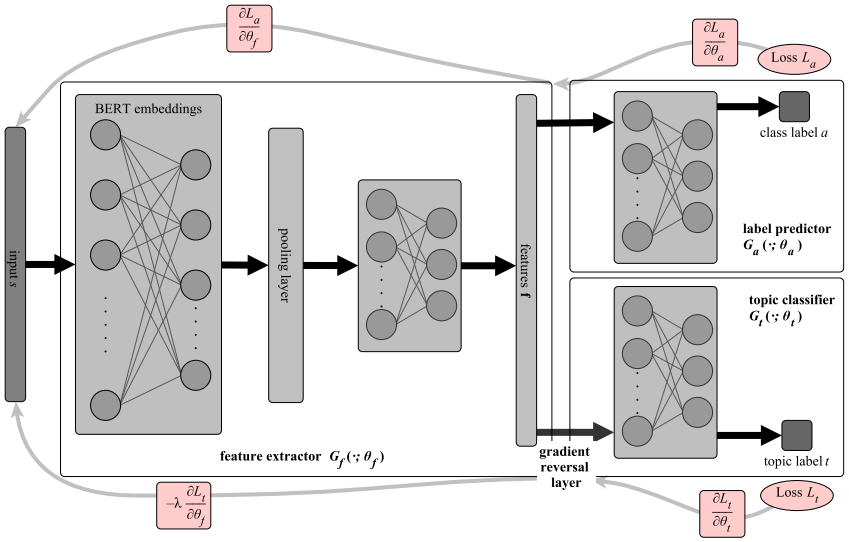

Figure 5: Gradient reversal procedure for debiasing topic information adapted from Ganin et al. (2015).

only, multiple binary classifiers are trained, one for each emotion in the source dataset. These binary classifiers are then applied to the target corpus, and only the most confident prediction is considered for the evaluation. Classification in the opposite case (single-label to multi-label) is achieved by training multiple binary classifiers, where each predicted label is considered in the evaluation. For any of the cross-domain classifications, the entire source corpus is used for training, and the entire target corpus is used as the testset.

3.4 Data

In the following, the 6 corpora considered in the inverstigation of topic bias are presented in more detail. Besides for their widespread use, the corpora are specifically selected for their variety in domain and text style. As bias in general and topic bias in particular is closely related to the respective dataset's domain, annotation and sampling methods of a dataset, the following overview puts emphasis on these aspects. An overview over general corpus statistics is provided in Table 5.

Corpus	Size	Annotation	Domain	Class. Setting
ISEAR	7666	Ekman + *shame*, *guilt*	event descr.	single
SSEC	4870	Plutchik	tweets	multi
TALES	10339	Ekman + *no emotion*	fairy tales	single
CROWD-	6600	Ekman + *shame*, *pride*, *bored.*, *rel.*, *trust*, *guilt*, *no*	event descr.	single
ENVENT		21 appraisal dimensions		multi
APPREDDIT	780	*unexp.*, *consist.*, *cert.*, *cntrl.*, *resp.*	reddit posts	multi
ENISEAR	1001	*attent.*, *cntrl.*, *circum.*, *resp.*, *pleasant.*, *effrt.*, *cert.*	event descr.	multi

Table 4: Overview over number of instances (size), annotation scheme, domain and classification setting (single-label or multi-label). For CROWD-ENVENT, the statistics for the emotion and appraisal annotations are reported separately

ISEAR The ISEAR corpus (Scherer and Wallbott, 1994) consists of 7,665 sentences which were sampled in a crowd-sourcing setup: Participants were presented with an emotion label and asked to report an event that elicited that particular emotion in them. Each event description is labeled with a single emotion from a set of eight (Ekman's basic emotions plus *shame* and *guilt*). Since participants were free to report any event that elicited one of the given emotions, they were also free in their choice of topic. However, since participants were asked to report events specific to certain emotions, sample bias could have been introduced to the corpus (under the assumption that there are prototypical event for certain emotions).

SSEC The Stance Sentiment Emotion Corpus (Schuff et al., 2017) consists of 4,868 Twitter posts. The original data stems from (Mohammad et al., 2016) which Schuff et al. (2017) re-annotate for emotions (Plutchik) by expert annotators. The annotations are conducted by trained expert annotators. Since the original dataset by Mohammad et al. (2016) was developed for stance detection, the isntances were sampled using keywords (i.e., hashtags) that contain a particular stance in favor (e.g., "#Hillary4President"; p. 33) or against an entity ("#HillNo"). This type of keyword-based data sampling has been related to topic bias in related investigation, e.g., on datasets of abusive language (Wiegand et al., 2019; cf. 2.4.1).

TALES The Tales corpus (Alm et al., 2005) features 15,302 sentences from different fairytales. Sentences are labeled by experts with one of Ekman's basic emotions (*surprise* is split into *negative* and *positive surprise*). Emotions are annotated from the perspective of the respective character.

CROWD-ENVENT Analogous to ISEAR, the crowd-enVENT corpus (Troiano et al., 2023) consists of 6600 crowd-sourced, self-reported event descriptions. Each description is annotated for 21 appraisal dimensions[12], each rated on a scale between 1 and 5, as well as for emotions (Ekman's 6 basic emotions, plus *shame*, *pride*, *boredom*, *relief*, *trust*, *shame*, *guilt* and *no emotion*). Participants were free in their choice of topic, but the priming with an emotion label might influence of topic distribution (see ISEAR). In order to avoid oversampling descriptions of prototypical events, Troiano et al. (2019) apply a diversification method to foster more diverse event descriptions. The corpus additionally features crowd-sourced re-annotations of the event descriptions to investigate differences between the reader's and writer's assessment of emotions and appraisals. However, these are excluded here.

APPREDDIT. The APPReddit corpus (Stranisci et al., 2022) is annotated with appraisal dimensions. It comprises 780 reddit posts, where each posts contains at least one event description (1,091 events overall). The five appraisal labels (*certainty*, *consistency*, *control*, *unexpectedness*, *responsibility*) are based on **?** and annotated by experts. The posts are sampled exclusively from a limited set of subreddits, mostly connotated with negative sentiment (Anger, offmychest, helpmecope anxiety, i.a.). This sampling procedure might introduce bias to the dataset.

ENISEAR The corpus consist of 1001 event descriptions that were originally compiled by Troiano et al. (2019) as a complement to ISEAR. However, in the context of this work,

[12]Suddeness, familiarity, event predictability, pleasantness, unpleasantness, goal relevance, own responsibility, others' responsibility, situational responsibility, anticipation of consequences, goal support, urgency, own control, others' control, situational control, acceptance of consequences, clash with internal standards and ideals, violation of (external) norms and laws, not consider, attention, effort.

Corpus	A	D	F	J	Sa	Sh	Su	No	O
CROWD-ENVENT	550	550	550	550	550	550*	550	550	2,200*
ISEAR	1,096	1,096	1,095	1,094	1,096	2,189*	–	–	–
SSEC	1388	440	274	815	414	–	177	1552	1077*
TALES	302	40	251	579	340	–	144	8,683	–

Table 5: Number of instances of each emotion class (after mapping; the asterisk (*) indicates that this class includes mapped labels, i.e., combining multiple classes into one aggregated, but not simple one-to-one mapping (happiness -> joy)).

ENISEAR refers to the appraisal annotations added to the corpus by Hofmann et al. (2020): *Attention*, *certainty*, *effort*, *pleasantness*, *responsibility* and *control*. The corpus has been annotated by three expert annotators.

3.4.1 Aggregated Annotation Scheme

As depicted above, the corpora differ significantly in their annotation schemes. In order to provide a more comparable analysis, the individual annotations are mapped onto a inter-corpora annotation scheme. For emotions, *anger*, *disgust*, *fear*, *joy*, *sadness*, *shame*, *surprise*, *no emotion* and *other* are considered. This subset of emotion labels is mainly based on the set of basic emotions proposed by Ekman (1999). Beyond Ekman's six emotions, the list accounts for other labels that frequently occur in the corpora (see Table 5 for an overview). The same procedure is applied to appraisal labels. However, approaches to appraisal classification are even more diverse in annotation than emotion datasets. To account for this variation, the inter-corpora labelset consists of 11 appraisal dimensions (suddenness, pleasantness, self control, chance control, self responsibility, other responsibility, goal support, predict consequences, attention, effort), however, only a subset of six labels is shared across two of the three corpora annotated with appraisals, while only two labels can be mapped to all three corpora (summarized in Table 6).

Corpus	Attention	Pleasantness	Suddenness	Self Control	Chance Control	Self Responsibility	Other Responsibility	Predict Consequences	Goal Support	Effort	Other
APPREDDIT	–	–	307	307	–	400	457	748	312	–	–
CROWD-ENVENT	4125	2261	3128	2142	1514	2597	3396	2841	2281	3210	6527*
ENISEAR	673	149	–	228	240	377	–	761	–	400	–

Table 6: Number of instances of each appraisal class (after mapping; the asterisk (*) indicates that this class includes mapped labels, either by simple one-to-one mapping (happiness -> joy), or by combining multiple classes into one aggregated).

4 Results

4.1 Topics in Emotion Datasets

Table 7 reports the results of the topic modeling at the overall corpus level. The overview includes the number of topics, the average size (number of instances) and the list of topic labels (L_c) fo reach corpus. These are defined manually, based on the ten most representative words for each topic (for the full lists of topic words, refer to Appendix B).

4.1.1 Topic Distribution

The analysis can be approached from two perspectives: First, by investigating how the topics within each corpus are related, and second, how the topics compare across corpora. In the following, the distribution of topics (as generated by the topic model) are discussed along these two perspectives.

	# Topics	∅ Topic	StD	Topic labels	Corpus size	*Outlier*
ISEAR	10	525	290	love, exams, death, shame, school, animals, alcohol, accidents, fear, theft	7666	2412
SSEC	11	305	219	feminism, prayer, abortion, climate, clinton, twitter, trump, gay marriage, latino, swearing, patriotism	4870	1513
TALES	10	388	183	birds, flowers, tabitha twitchit, old english, piggies, royalty, dressmaking, hansel & gretel, boats, predators	10339	6457
CROWD-ENVENT	8	584	298	feelings, promotion, relationships, covid, dogs, graduation, pregnancy, driving	6600	1925
APPREDDIT	10	43	12	depression, everyday life, driving, love, romantic relationships, reddit, anger, death, platonic relationships, vaccination	780	352
ENISEAR	13	58	25	death, dogs, accidents, theft, birth, food, affairs, UK politics, christmas, bullying, work, relationships, spooky	1001	245

Table 7: Number (#), average size (∅), standard deviation (StD), and the manual label of the topics found by BERTopic for all corpora. All numbers exclude the outlier class, whose size is given last (*Outlier*). The topic labels are sorted by size, in decreasing order. The second to last column reports the number of all instances per corpus for reference.

manual label	top 4 topic words	manual label	top 4 topic words
birds	bird, sang, birds, singing	royalty	princesses, princess, queen, princes
flowers	flowers, blossoms, flower, blossom	dressmaking	tailor, dressed, garments, shop
tabitha twitchit	duchess, ribby, tabitha, kitten	hansel & gretel	hans, hansel, carpenter, shepherd
old english	thou, thee, thy, mercy	boat	boat, sail, sailed, sailing
piggies	pigling, pig, piggies, pigs	predators	fox, foxy, wolf, lion

Table 8: Topics in TALES by manually defined label and 4 most representative topic words.

Within corpora Investigating the topic distribution within each corpus (Table 7) shows that the individual corpus' domain and sampling methods are reflected in the topic. ISEAR, ENISEAR and CROWD-ENVENT, all of which are compiled by querying emotionally connotated event-descriptions, feature generic and everyday topics, e.g., *love*, *dogs* or*driving*. In SSEC, on the other hand, topics are very specific to the keyword-based sampling method applied by Mohammad et al. (2016) to compile the corpus (cf. 3.4): The topics revolve around twitter debates concerning climate change (topic *climate*) or US politics (*clinton*, *trump*). In APPREDDIT, topics appear to be indicative of the subreddit they are sourced from. For instance, the topic of *depression* could be related to the subreddit "mentalhealth" (or, similarly, "helpmecope" or "anxiety"), and the variety of relationship-related topics (*romantic relationships*, *love*, *platonic relationships*) appears to reflect the various subreddits revolving around these topics, e.g., "relationship advice" or "Dear Ex" (cf. Stranisci et al., 2022 for the exhaustive list of sampled subreddits). Compared to the other corpora, the topics in TALES appear to be the noisiest in terms of inter-topic homogeneity. Table 8 reports the 4 most representative words for each topic in order to provide more insight on topics in TALES are constituted. From this, it can be observed that some topics are defined by semantic relatedness (*birds*, *flowers*, *royalty*), while others are representative of specific fairy tales. For example, the topic labeled as *Tabitha Twitchit* comprises the names of fictional characters from the kids stories by Beatrix Potter. Further, the topic *old english* appears to be based on lexical features alone. TALES thus comprises a mix of literally, semantically and contextually defined topics.

A further observation related to the domains of the respective corpora is that APPRED-

Corpus	Topics related to *Driving/ Accident*	Topics related to *Relationships/Love*
ISEAR	accident, driving, bus, drove	love, lover, relationship, loved
CROWD-ENVENT	accident, driving, drove, car	married, friendship, fiance, relationship
APPREDDIT	restaurant, driving, radio, taco	love, loving, loves, loved
APPREDDIT		breakup, gf, girlfriend, relationship
APPREDDIT		friends, classmates, friendzone, socialized
ENISEAR	accident, hurt, driving, drove	friends, unkind, betrayed, meet

Table 9: Example of two broader topics occurring in four different corpora with most informative keywords showing the differences in topic focus across corpora.

DIT and SSEC both feature a "meta" topic, i.e., a topic that explicitly relates to the platform the instances are sampled from.(*twitter* and *reddit*, respectively).

Concerning the size of topics, i.e., the number of instances associated with it, a notable variation can be observed within all the investigated corpora (as indicated by the standard deviation of topic size). The corpora with the largest topics on average are CROWD-ENVENT and ISEAR, which are both comparable in overall size as well. Accordingly, the corpus with the least instances (APPREDDIT) also exhibits the smallest average topic size. In TALES, the average topic size is lower than in ISEAR and CROWD-ENVENT, despite TALES being significantly larger. However, the standard deviation of topic size in TALES is also lower.

Across corpora Across all corpora, the number of topics is quite consistent, ranging from 8 (CROWD-ENVENT) to 13 (ENISEAR), while ISEAR, TALES and APPREDDIT comprise 10, ISEAR 11 topics. As described above, the topics are specific to the respective corpus' domain. Therefore, the granularity of topics differs across corpora: Topics such as *love*, *death* (ISEAR) or *flowers* (TALES) are more coarse-grained and generic, while *gay marriage* (SSEC) or *vaccination* (APPREDDIT) are more fine-grained, thus reflecting the more confined domain space of the corpus. Despite their domain-specificity, some topics also occur across multiple corpora. For example, the closely related topics of *love*

and *relationship* appear across ISEAR, CROWD-ENVENT, APPREDDIT (which comprises multiple topics for distinct types of relationships) and ENISEAR. Similarly, *driving* as well as the related topic of *accident* are not exclusive to only on corpus. To give an overview over the nuances between superficially similar topics across corpora, Table 9 lists the topic words associated with closely related topics from different corpora. While the topics of *driving* or *accident* are very similar (based on the overlap of most representative words) accross ISEAR, CROWD-ENVENT and ENISEAR, the topic is slightly differently connotated in APPREDDIT. For the topic of *relationship* and *love*, the differences are more pronounced. In ISEAR, the topic is explicitly defined by romantic cue words, while in CROWD-ENVENT, the topic is a more coarse-grained representation of *relationship*, also comprising friendship. APPREDDIT, as stated above, features very nuanced topics of different types of relationships.

4.1.2 Topics and Emotions

One key assumption made in this work is that topics and emotions are related. One obvious relation is that some topics are defined through explicit emotion words, namely as *fear* (most represented by "frightened", "fear", "terrified", "fearful") and *shame* ("ashamed", "guilt", "shame", "depressed") in ISEAR. Further, CROWD-ENVENT comprises the topic of *feelings*, which, judging only by the most representative words ("feel", "feeling", "feelings", "pass"), is not associated with a specific emotion.

In order to assess whether these equivalences on the lexical are also present in the actual emotion annotations, Figure 6 reports the normalized pointwise mutual information (PMI) between topics and the associated emotion annotations in ISEAR. What can be observed is that the topic of *shame* is, indeed, highly positively correlated with the emotion label *shame*. Similarly, topic and emotion *fear* are highly correlated as well. Besides the obvious correlations between emotion and equivalent topics, further emotionally correlated topics are *death* (with *sadness*), *alcohol* and *animals* (both *disgust*), *accidents* (*fear*) and *exams* with *joy* (all positive). Strong negative correlations can be observed for *alcohol* and *joy*, as well as for *love* and *fear*. Table 10 illustrates these correlations through some example sentences for highly correlated emotion-topic pairs in ISEAR. Based on the reported

Figure 6: Normalized pointwise mutual information between topics and emotion annotations in ISEAR.

Emotion	Topic	Text
sadness	death	'When my mother died.'
disgust	alcohol	'A friend of mine came to see me when he was quite drunk.'
joy	exams	'Passing an exam I did not expect to pass.'
fear	fear	'I felt fear when I was young and left in my big house all alone.'

Table 10: Prototypical sentences for highly correlated topic-emotion pairs in ISEAR.

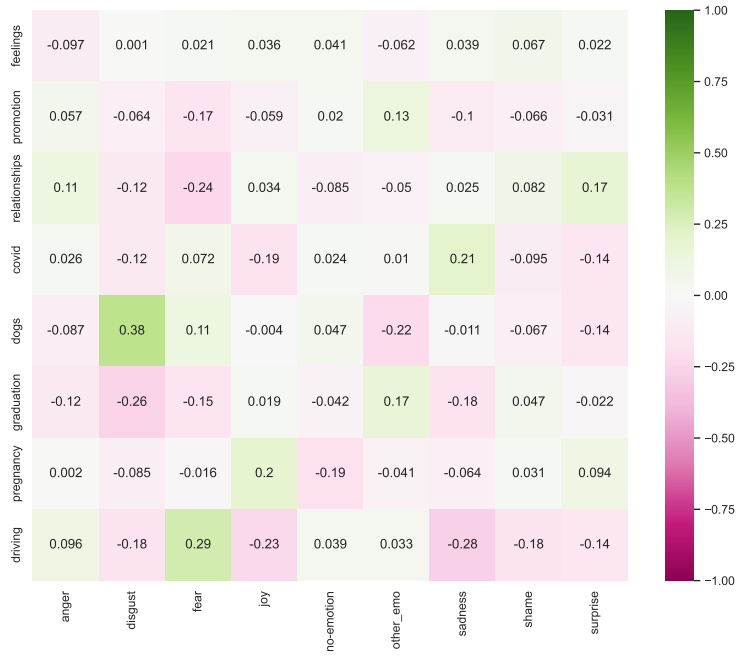

Figure 7: Normalized pointwise mutual information between topics and emotion annotations in CROWD-ENVENT.

correlations, there appear to be prototypical emotions for certain topics in ISEAR (which also makes sense intuitively). Similar observations can be made for CROWD-ENVENT, where a high positive correlation between *dogs* and *disgust*, as well as between *driving* and *fear* can be observed (Figure 7). Although these are consistent with correlations of similar topics in ISEAR (*animals* and *disgust*, *accidents* and *fear*), the PMI values in CROWD-ENVENT are consistently lower: The highest positive correlation in ISEAR is associated with a PMI of 0.6 (*fear*), while in CROWD-ENVENT, the strongest correlation (*dogs*, *disgust*) has a PMI of 0.38. For TALES and SSEC, the topic-emotion correlations are even less distinct (cf. Figure 18 and 17 in Appendix C).

As CROWD-ENVENT features annoations for both emotions and appraisals, is can be used to investigate how correlations between appraisals and topics differ from the correlations reported for topics and emotions (Figure 8). The direct comparison shows, that the correlations between topics and appraisals are significantly less distinct than for emotions (in CROWD-ENVENT). The highest positive correlation (PMI of 0.19) is between

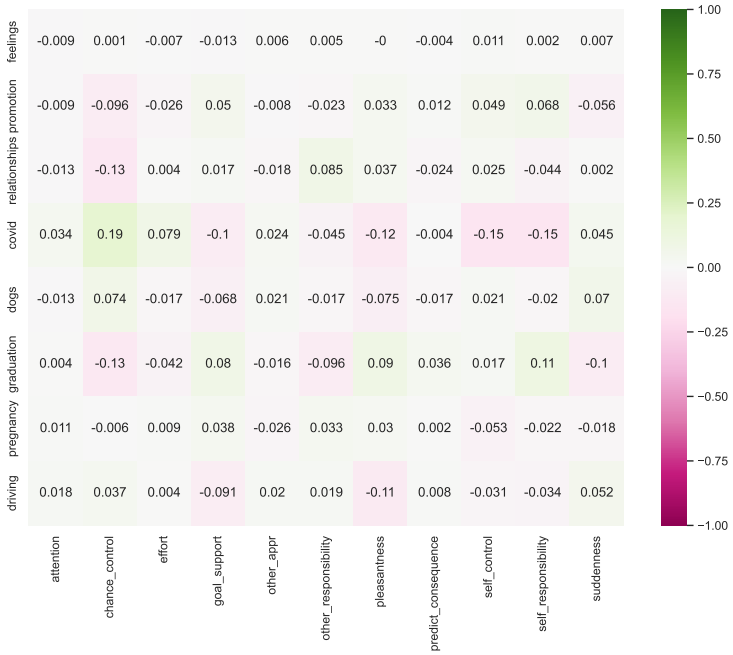

Figure 8: Normalized pointwise mutual information between topics and appraisal annotations in CROWD-ENVENT.

covid and *chance control*, i.e., *covid*-related events are appraised as out of control by the emoter. The topic of *covid* is further (slightly) negatively correlated with *self control* (thus, the complement to *chance control*) and *self responsibility*. PMI correlations for all other corpora are provided in Appendix C.

This qualitative analysis can serve as a starting point for the discussion of the subsequent, quantitative experiments.

4.2 Topic Bias in Emotion Classifiers

What arises from the observation that topics and emotions (i.e., topics and appraisals) are indeed correlated is the question whether this relation is reflected in emotion and appraisal classifiers.

To this end, the results of CROSSTOPIC and INTOPIC experiments are presented in Table 11. Columns CROSSTOPIC reports the average F_1 over all t_c^{out} in c . The F_1 reported in column INTOPIC is the average over all $\overline{t_{c,x}^{in}}$, where $\overline{t_{c,x}^{in}}$ is defined as the average over all $t_{c,x}^{in}$ for a given $t_c^{out} = x$.

Following the assumption that emotions and appraisals classifiers are biased towards topics, the INTOPIC setting is hypothesized to score higher than the CROSSTOPIC setting, indicated by a positive difference (Δ) between both settings. In fact, all corpora score higher in the INTOPIC setting than in CROSSTOPIC. However, the difference varies between corpora. The highest Delta is observed for ISEAR (+9), while the improvement in INTOPIC is only marginal (+1) for SSEC, TALES and CROWD-ENVENT (in the appraisal classification setting) and APPREDDIT (+2). In comparison, CROWD-ENVENT (for emotion classification) and ENISEAR show moderate improvement when evaluated INTOPIC (+4 and +5, respectively). Overall, the Delta values are similar (on average) between emotion and appraisal classification.

In order to enable a more detailed analysis, Figure 9 reports the F_1-scores obtained on each topic-specific subset t_c^{in} and t_c^{out} for each held-out topic. The diagonal thus depicts the CROSSTOPIC setting (INTOPIC for all other) for ISEAR. The high Delta reported for ISEAR can also be observed in Figure 9 as the F_1-scores in the diagonal (CROSSTOPIC) are

	Corpus	Average F_1 CROSSTOPIC	INTOPIC	Δ
Emotion	ISEAR	59	68	9
	SSEC	46	47	1
	TALES	84	85	1
	CROWD-ENVENT	51	55	4
	Average	60	64	4
Appraisal	CROWD-ENVENT	63	64	1
	APPREDDIT	66	68	2
	ENISEAR	70	75	5
	Average	66	69	3

Table 11: Results for the cross-topic experiments on emotion and appraisal corpora. Shown are micro-F_1-scores averaged over all evaluations on unseen topics (CROSSTOPIC), all evaluations on topics seen in training (INTOPIC) and the difference between the two (Δ), where positive values denote a better in-topic performance.

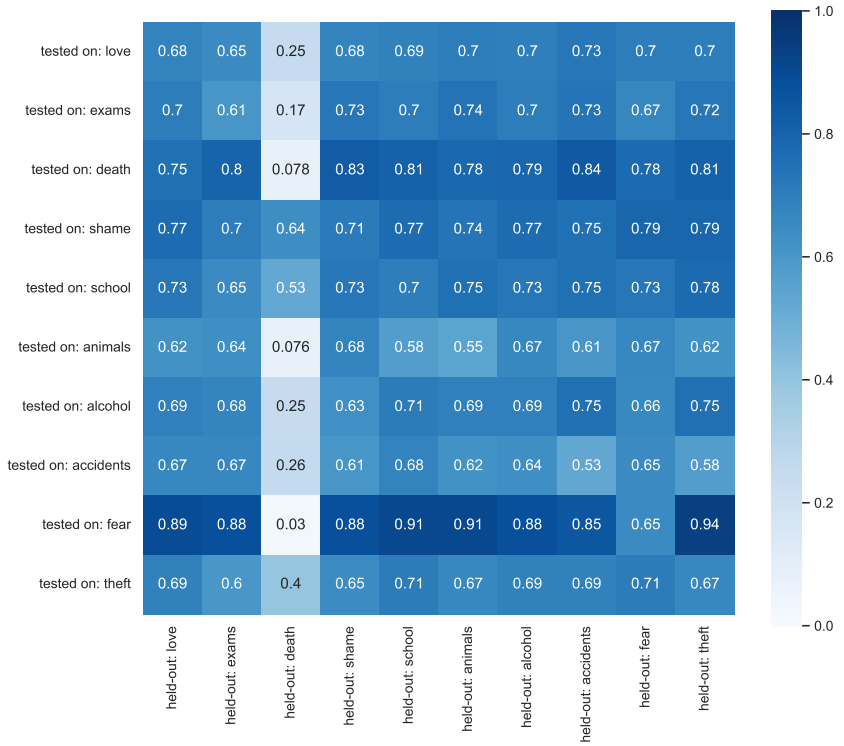

Figure 9: Micro-average F_1 for each testset in ISEAR, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

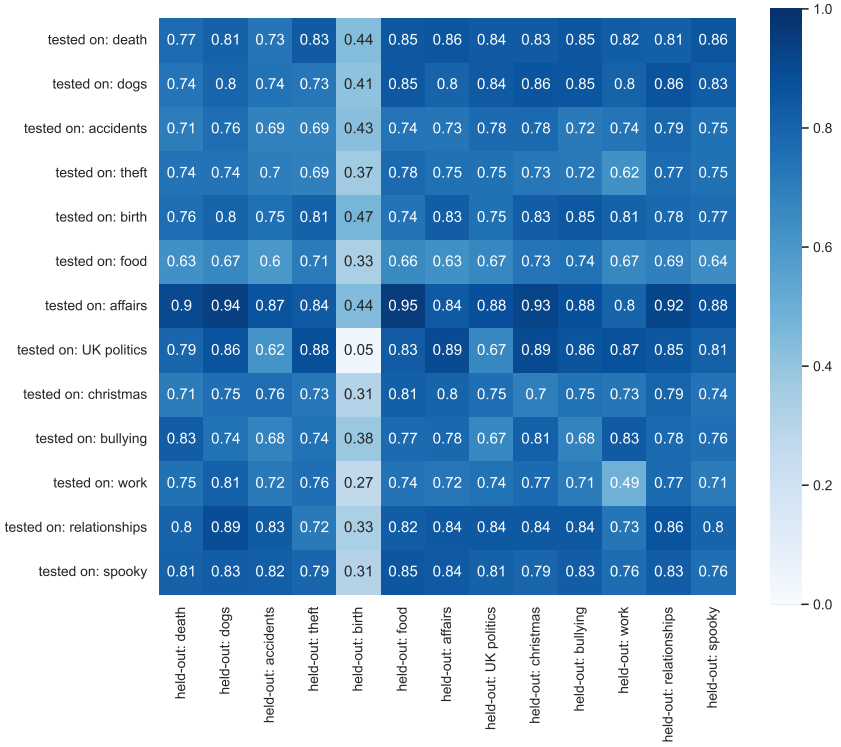

Figure 10: Micro-average F_1 for each testset in ENISEAR, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

consistently lower than the average of all other results of the same held-out topic (INTOPIC). That being said, the CROSSTOPIC scores are still comparably high. What can further be inferred is that when the topic of *death* is absent from the training data, the classifier performs much worse on all testsets, both INTOPIC and CROSSTOPIC. Analogously, the topic *fear* appears to be easier to classify, no matter which held-out topic is absent from the training data. The only exception is the mentioned topic *fear*, and, although to a lesser extent, in the CROSSTOPIC setting of topic *fear*.

Another corpus that shows a significant Delta between CROSSTOPIC and INTOPIC is ENISEAR (Figure 10). Similar to ISEAR, the decrease in CROSSTOPIC performance can be observed in the diagonal. In ENISEAR, the topic of *birth* has a similar effect on classification performance as *death* in ISEAR, in the sense that performance decreases across all testing susbsets if that topic is not among the training data.

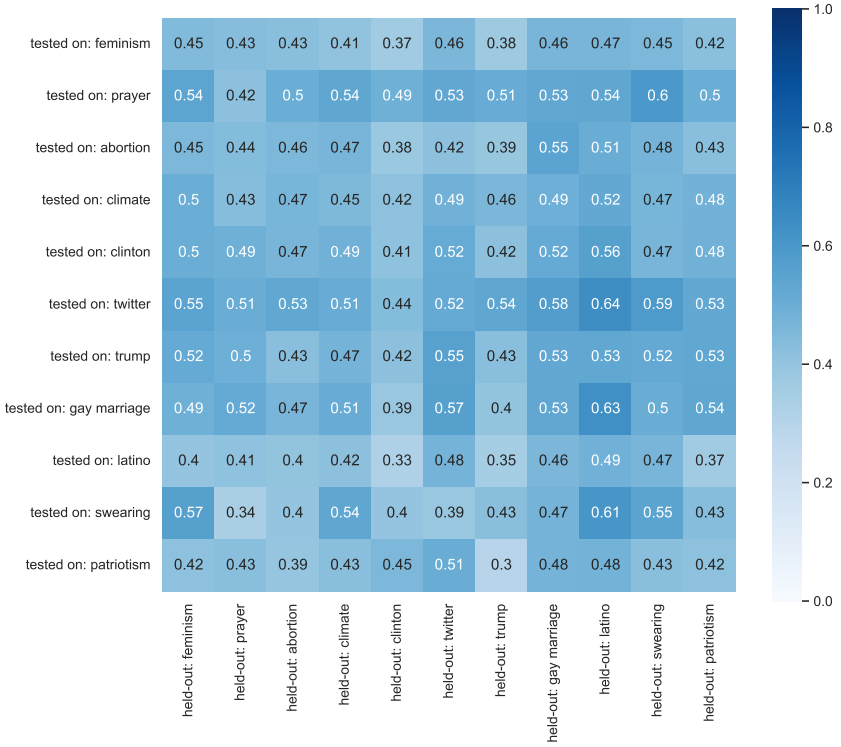

Figure 11: Micro-average F_1 for each testset in SSEC, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

However, for SSEC, no significant difference in F_1 can be observed between the two settings. From the depiction in Figure 11, three insights are to be gained: First, the overall classification performance on SSEC is worse than on ISEAR and enISEAR. Second, the F_1-scores in the diagonal are not (consistently) lower than the average scores achieved for one held-out topic. Third, the results obtained across testsets (both InTopic and CrossTopic) are lower when certain topics are excluded from the training data, namely the topics *clinton* and *trump*.

4.3 Topic Bias mitigation

Given the reported decrease in performance in CrossTopic (as compared to InTopic), the following experiments investigate whether these topic-specific differences can be mitigated.

4.3.1 Word Removal

First, mitigation via word removal (InTopic-mask and CrossTopic-mask) is investigated. The results are presented in Table 12, analogous to Table 11.

Assuming that word removal proves effective for mitigating topic bias, the Delta between CrossTopic-mask and InTopic-mask should decrease in comparison to the standard classifier in 11. However, this is generally not the case: For ISEAR, the Delta further increases, due to a slight increase in InTopic-mask performance. For crowd-enVENT and Tales, no change in performance in either direction can be observed. On SSEC, the overall perfomance drops significantly, from 46 F_1 in CrossTopic and 47 F_1 in InTopic to 37 F_1 and 39 F_1 in CrossTopic-mask and InTopic-mask, respectively. The inital Delta of +1 thus slightly increases to +2. Concerning the appraisal corpora, the small delta of +1 in crowd-enVENT is mitigated (Delta 0) as both CrossTopic-mask and InTopic-mask performance slightly decrease. APPReddit performs worse than in InTopic and CrossTopic as well, however, InTopic-mask performance drops even below CrossTopic-mask, resulting in a negative Delta. Only for enISEAR, a significant mitigation effect can be observed (from Delta +5 to +1), due to a stronger decrease for InTopic-mask than for

		Average F_1 with Word Deletion		
	Corpus	CROSSTOPIC-MASK	INTOPIC-MASK	Δ
Emotion	ISEAR	59	70	11
	SSEC	37	39	2
	TALES	84	85	1
	CROWD-ENVENT	51	55	4
	Average	57	62	5
Appraisal	CROWD-ENVENT	61	61	0
	APPREDDIT	56	55	-1
	ENISEAR	56	57	1
	Average	57	57	0

Table 12: Results for the cross-topic word deletion experiments. The information shown here is analogous to that of Table 11 but with the added debiasing method of masking the ten most informative words of each topic with "…" in both training and test data (CROSSTOPIC-MASK/INTOPIC-MASK).

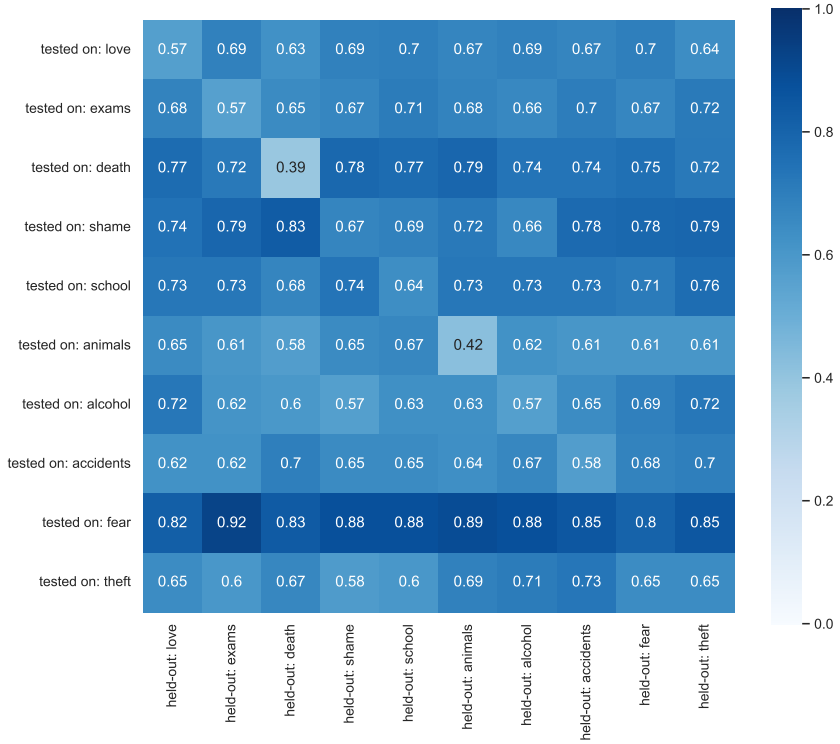

Figure 12: Micro-average F_1 for each testset in ISEAR, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

CROSSTOPIC-MASK (although both settings experience a significant performance decrease compared to INTOPIC/CROSSTOPIC). The described effects for ISEAR can also be observed in Figure 12. As reported in 12, there is no notable decrease in performance compared to the standard classifier in CROSSTOPIC and INTOPIC (cf. Figure 9). However, what has changed, is that the strong impact of topic *death* (when absent from the training data) is mitigated. The increased performance when testing on topic *fear* still persists.

Figure 13 reports the detailed results for ENISEAR, which are overall lower than for in the standard setting without mitigation method applied. As the word removal mitigation method proved successful in the case of ENISEAR, the diagonal in 13 is not significantly distinct from the other reported F_1 scores (as opposed to the results obtained using the non-mitigated classifier). Further, the drastic performance decrease caused by the absence of *birth* from the training data is mitigated as well, similar to the effect of word removal on

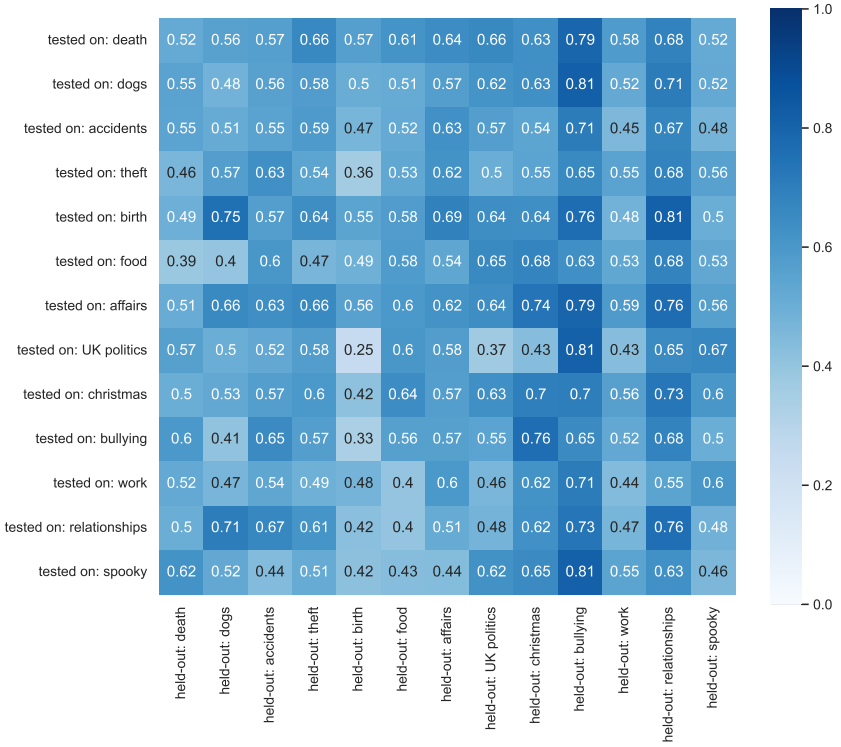

Figure 13: Micro-average F_1 for each testset in ENISEAR, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

	Corpus	Average F_1 with GRL		
		CROSSTOPIC-GRL	INTOPIC-GRL	Δ
Emotion	ISEAR	65	71	6
	SSEC	23	25	2
	TALES	82	83	1
	CROWD-ENVENT	54	57	3
	Average	56	59	3
Appraisal	CROWD-ENVENT	44	45	1
	APPREDDIT	56	56	0
	ENISEAR	54	56	2
	Average	51	52	1

Table 13: Results for the cross-topic gradient reversal experiments. The information shown here is identical to that of Table 11 but with the added debiasing technique of a second classification head for topics, whose gradient is reversed during training to debias the neural net against topic features.

the topic of *death* in ISEAR.

4.3.2 Gradient Reversal Layer

The results of the alternative approach to topic bias mitigation, gradient reversal, are reported in Table 13.

Compared to topic word removal, GRL appears to be the more capable bias mitigation method. For ISEAR, the original Delta of +9 is mitigated to +6, which is due to an increase in CROSSTOPIC-GRL (as compared to CROSSTOPIC). For CROWD-ENVENT, the Delta is decreased as well, although only slightly (+4 to +3). However, there is no change in Delta for TALES, and is even slightly increased for SSEC (from +1 to +2). Across all emotion corpora, the Delta is slightly reduced from originally, i.e., not mitigated +4 to +3 using GRL. However, this overall mitigation effect is mostly due to the decrease of Delta in ISEAR.

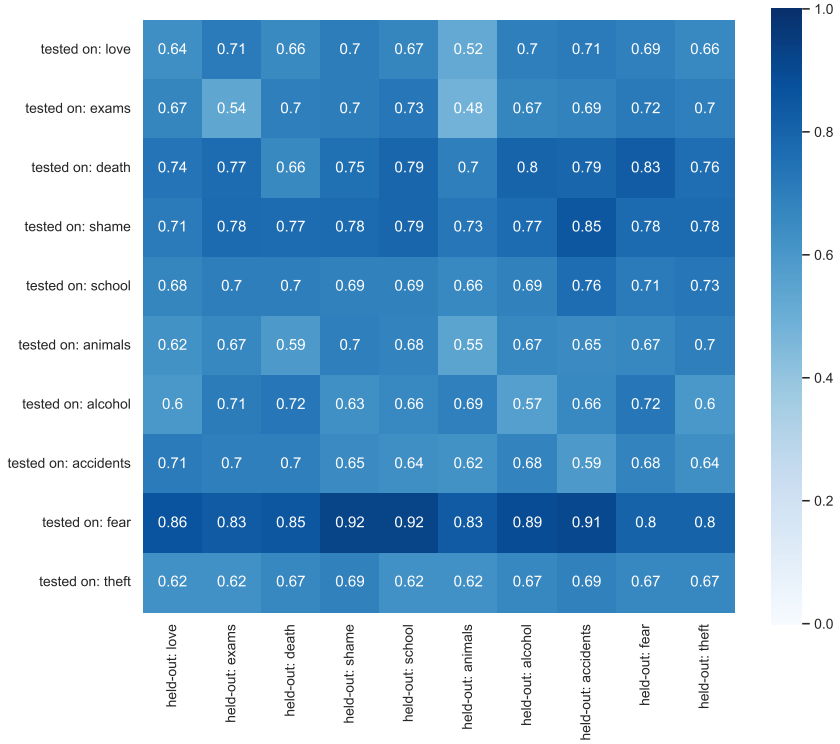

Figure 14: Micro-average F_1 for each testset in ISEAR, in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

For the corpora annotated with appraisal, a mitigation effect can be observed for both APPREDDIT (+5 to 2) and ENISEAR (+2 to 0), while the Delta of CROWD-ENVENT remains unchanged. Overall, the Delta drops from +3 to +1 with GRL.

Investigating the detailed results for ISEAR (Figure 14) shows again (as for mitigation using word removal) that the original impact of topic *death* is mitigated. Again, classifiers yield best results across all testsets when evaluated on topic *fear*.

5 Discussion

Table 14 summarizes the results for all CROSSTOPIC and INTOPIC experiments. In the first, non-debiased experiment, INTOPIC performance is found to be higher than CROSSTOPIC

		CROSSTOPIC			INTOPIC			Δ		
	Corpus	BASE	WR	GRL	BASE	WR	GRL	BASE	WR	GRL
Emotion	ISEAR	59	59	65	68	70	71	9	11	6
	SSEC	46	37	23	47	39	25	1	2	2
	TALES	84	84	82	85	85	83	1	1	1
	CROWD-ENVENT	51	51	54	55	55	57	4	4	3
	Average	60	57	56	64	62	59	4	5	3
Appraisal	CROWD-ENVENT	63	61	44	64	61	45	1	0	1
	APPREDDIT	66	56	56	68	55	56	2	-1	0
	ENISEAR	70	56	54	75	57	56	5	1	2
	Average	66	57	51	69	57	52	3	0	1

Table 14: Results for CROSSTOPIC and INTOPIC experiments and differences between them for all experimental series. For each experiment setup, the results for the baseline without debiasing (BASE) and the two debiasing methods of word removal before (WR) and an added gradient reversal layer (GRL) during training.

performance. Although the trend holds across all corpora, there are corpus-specific differences. The corpora exhibiting the highest Delta between both settings are, in order of decreasing Delta, ISEAR (+9), ENISEAR (+5) and CROWD-ENVENT for emotions (+4). APPREDDIT only shows a slight Delta of +2 and for CROWD-ENVENT (for appraisals), SSEC and TALES, the difference is even smaller (+1). Thus, a significant impact of topic bias on emotion classification can only be inferred in the case of ISEAR, CROWD-ENVENT (emotion) and ENISEAR.

In the subsequent experiment, the most relevant topic words are removed in order to mitigate the impact of topics on the classification. However, this method does not prove effective. For most corpora, the difference in performance between CROSSTOPIC-MASK and INTOPIC-MASK is not, as intended, reduced. For ISEAR and SSEC, the Delta increases (+2 and +1 in Delta, respectively), while for APPREDDIT, the CROSSTOPIC-MASK setting even outperforms the INTOPIC-MASK setting (overall Delta of -1). No change at all on average is reported for TALES and CROWD-ENVENT (emotions), and only in the case of CROWD-ENVENT (appraisals) a very slight decrease in Delta is observed (from 1 to 0), a

more significant one for enISEAR (from +5 to +1). Overall, the effect of word removal as a method for topic bias mitigation is not consistent across corpora.

In contrast, if gradient reversal is applied for bias mitigation, the Delta between the CROSSTOPIC-GRL and INTOPIC-GRL decreases – compared to CROSSTOPIC and INTOPIC– for ISEAR (-3), CROWD-ENVENT (emotions, slight decrease of -1), ENISEAR (-2) and APPREDDIT (-2). For SSEC, however, the Delta slighty increases. For CROWD-ENVENT (appraisals) and TALES, the Delta remains unchanged.

The here presented results lead to the following observations: First **(1)**, the overall impact of topic bias on emotion and appraisal classification appears to be less pronounced as hypothesized. In any case, the substantially different results obtained on different corpora indicate that the initially formulated research questions cannot be answered for emotion classification in general, but with respect to a specific corpus. This leads to second **(2)**: Based on the performance differences between INTOPIC and CROSSTOPIC, ISEAR is the most affected by topic bias, followed by ENISEAR and CROWD-ENVENT (for emotion classification). On the other end of the spectrum, TALES appears to be the most unaffected by topic bias. The difference between CROSSTOPIC and INTOPIC is not significantly pronounced and none of the mitigation methods that were applied had an effect, neither on the Delta between settings, nor on the overall F_1: TALES is the only corpus that does not exhibit a (significant) change in F_1 across experimental settings. Third **(3)**, the effect of word removal for bias mitigation produces inconsistent results across corpora, leading to either an increase (ISEAR), decrease (ENISEAR) or reversal (APPREDDIT) of Delta. Gradient reversal, while also not consistent across all corpora (light increase for SSEC), appears to be the overall better choice for bias mitigation. Fourth **(4)**, concerning the differences between emotion and appraisal corpora with respect to the influence of topics, no general statement can be made. Based on the minimal pair that is CROWD-ENVENT, appraisal classification seems to be less affected by topics than emotion classification. However, ENISEAR shows a comparably pronounced performance decrease in CROSSTOPIC.

I address both **(1)** and **(2)** by referring to the qualitative analysis conducted prior to the quantitative experiments (4.1). The analysis shows that the topics generated by the topic model are, most prominently, representative of the underlying sampling strategies: For

ISEAR, CROWD-ENVENT and ENISEAR, topics consist of generic, mostly unrelated topics that were reported by participants when asked to recall and emotional event. For SSEC, topics revolve around the keywords, i.e., twitter hashtags, that were used for sampling them (.e.g, about Donald Trump or HIllary Clinton). TALES reflects, at least in parts, the stories and fairy tales the corpus is comprised of, e.g., Tabitha Twitchit. In APPREDDIT, the topic distribution follows the topics of the sampled subreddits. Based on this observation, I hypothesize that every emotion dataset investigated here is, in fact, biased towards topics, in the sense that the topics in each dataset reflect the underlying sampling method. By itself, this does not pose a problem to emotion classification. Only if the sampling method is biased in itself, that bias is propagated to the level of emotion or appraisal annotations. This hypothesis is grounded on the observation that ISEAR exhibits the most topic bias. This bias shows in the classifiers (i.e., in the difference between CROSSTOPIC and INTOPIC), as well as in the correlations between topics and emotion labels, as depicted in Figure 6. In ISEAR, the biased sampling method (i.e., querying event descriptions associated with a given emotion) leads to event descriptions which are prototypical for certain emotions – and, in consequence, to topics within the dataset which are prototypical for these emotions as well. This hypothesis does also account for the fact that both ENISEAR and CROWD-ENVENT, two corpora that applied the same sampling strategy, exhibit the most topic bias besides ISEAR. I further hypothesize, that the method applied by Troiano et al. (2019) in order to prevent sampling overly prototypical event descriptions (i.e., generating lists of undesired events) is the reason for why CROWD-ENVENT exhibits the least amount of bias among these three corpora and, further, why the calculated PMI does not reveal as prototypical topic-emotion relations as it does for ISEAR. However, this hypothesis is based on anecdotal evidence and needs further testing, i.e., by recreating the experiments conducted here for source-representations instead of topics.

Concerning the unexpected results for TALES, I assume this to be related to the over-proportionally large subset annotated as neutral, i.e., with *no emotion* (cf. Table 6). Even if TALES is biased towards topics, this large portion of data masks any effect of the experiments. Additionally, TALES is the noisiest corpus in terms of topics, as detailed in the analysis (4.1)

For **(3)**, i.e., the unpredictable effect of applying word removal for bias mitigation, I

hypothesize the variations in results to be caused by the fact that some words carry more relevant information for the classification process than others, regardless of whether these words are also representative of topics. I these are removed, the consequences can vary depending on the distribution of data in train- and testset. In this context, gradient reversal is more precise, as it implements a trade-off between classification performance an reliance on topics words.

(4) A decisive factor concerning the performance of the appraisal datasets in the experiments is the annoation scheme. While CROWD-ENVENT is comprised of a large number of appraisal dimensions (21), APPREDDIT and ENISEAR comprise significantly less. As a result, the classification task for CROWD-ENVENT is more complex and, in addition, the large number of labels hinders a simple prototypical association of topics and appraisal dimensions. For this reason,i.e., due to its smaller labelset, APPREDDIT appears to be more prone to topic bias than CROWD-ENVENT.

6 Conclusion

This work addressed a novel perspective on computational emotion analysis by investigating the impact of topics on the classification of emotions and appraisals.

The analysis of topic distribution in emotion corpora yields that available emotion corpora are, in part, biased towards topics. However, the degree of bias varies greatly. Some corpora exhibit prototypical topics for certain emotions, while in others, no relation between topic and emotion distribution can be observed. This variance is hypothesized to be related to the sampling strategy applied in aggregating the corpus: If the sampling method is biased, i.e., if certain topics are over-represented for a given emotion, topic bias emerges.

In the case that topic and emotion distribution are highly correlated, this topic bias is also reflected in resulting classifier. For mitigating this bias in emotion classifiers, two approaches were considered. As a naive approach, topic-specific words were completely removed from the corpus. However, this method results in a non-transparent change in performance. For some corpora this method achieves debiasing, but for most, it does

not. As an alternative, gradient reversal is applied, in order to mitigate the influence of topic-specific features on the classification task. This method proved more successful and more reliable than word removal.

This work bridges previous research on topic bias with emotion and appraisal classification and contributes towards the development of more robust classifiers.

A one potential avenue for future work is to investigate to which degree the corpus-specific topic distributions are contributing to the challenge of domain-specificity in emotion classification. If topics are found to hinder cross-domain and cross-corpus classification, topic bias mitigation methods could be applied in order to assess their potential for domain-adaptation.

References

Ahmed Abbasi and Hsinchun Chen. Affect intensity analysis of dark web forums. In *2007 IEEE Intelligence and Security Informatics*, NJ, USA, 2007.

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121, 2020. doi: 10.1109/ICCWAMTIP51612.2020.9317379.

Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 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, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1073.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. 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 1941–1955, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.151. URL https://aclanthology.org/2021.acl-long.151.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3805. URL https://aclanthology.org/W19-3805.

C. D. Batson, Laura L. Shaw, and Kathryn C. Oleson. Differentiating Affect, Mood, and Emotion: Toward Functionally-Based Conceptual Distinctions. In M. S. Clark, editor,

Review of Personality and Social Psychology, volume 13, pages 294–326, Newbury Park, CA, 1992. Sage.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.nlperspectives-1.2.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. ISSN 1532-4435.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-1615.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1056.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Laura-Ana-Maria Bostan and Roman Klinger. 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, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1179.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.194.

Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1052.

Sven Buechel and Udo Hahn. 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, April 2017a. Association for Computational Linguistics. URL https://aclanthology.org/E17-2092.

Sven Buechel and Udo Hahn. 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, April 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-0801. URL https://aclanthology.org/W17-0801.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ltedi-1.11. URL https://aclanthology.org/2022.ltedi-1.11.

Yung-Chun Chang, Cen-Chieh Chen, Yu-Lun Hsieh, Chien Chin Chen, and Wen-Lian Hsu. Linguistic template extraction for recognizing reader-emotion and emotional resonance writing assistance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 775–780, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2127. URL https://aclanthology.org/P15-2127.

Annie Chen. Patient experience in online support forums: Modeling interpersonal interactions and medication use. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 16–22, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-3003.

Arjun Choudhry, Inder Khatri, Arkajyoti Chakraborty, Dinesh Vishwakarma, and Mukesh Prasad. Emotion-guided cross-domain fake news detection using adversarial domain adaptation. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 75–79, New Delhi, India, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.icon-main.10.

Alexandra Ciobotaru and Liviu P. Dinu. RED: A novel dataset for Romanian emotion detection from tweets. In *Proceedings of the International Conference on Recent Advances*

in Natural Language Processing (RANLP 2021), pages 291–300, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.34.

W. Cooke. *Mind and the emotions, considered in relation to health and disease*. Longman, 1838.

Kate Crawford. The trouble with bias. keynote at neural information processing systems (NIPS'17).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3504. URL https://aclanthology.org/W19-3504.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.24.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. Semi-supervised topic modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Alvin Dey, Tanya Chowdhury, Yash Kumar, and Tanmoy Chakraborty. Corpora evaluation and system bias detection in multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2830–2840, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.254. URL https://aclanthology.org/2020.findings-emnlp.254.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173986. URL https://doi.org/10.1145/3173574.3173986.

Thomas Dixon. "emotion": One word, many concepts. *Emotion Review*, 4(4):387–388, 2012. doi: 10.1177/1754073912445826. URL https://doi.org/10.1177/1754073912445826.

Yuchang Dong and Xueqiang Zeng. Lexicon-enhanced multi-task convolutional neural network for emotion distribution learning. *Axioms*, 11(4), 2022. ISSN 2075-1680. doi: 10.3390/axioms11040181. URL https://www.mdpi.com/2075-1680/11/4/181.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.370. URL https://aclanthology.org/2020.acl-main.370.

Anton Eklund and Mona Forsman. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-industry.65.

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6:169–200, 1992.

Paul Ekman. *Basic Emotions*, chapter 3, pages 45–60. John Wiley & Sons, Ltd, 1999. ISBN 9780470013496. doi: https://doi.org/10.1002/0470013494.ch3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch3.

Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370, 2011. doi: 10.1177/1754073911410740. URL https://doi.org/10.1177/1754073911410740.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL https://aclanthology.org/D18-1002.

Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf.

Lisa Feldman Barrett and James A. Russell. Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, 74:967–984, 1998.

Oliver Ferschke, Iryna Gurevych, and Marc Rittberger. The impact of topic bias on quality flaw prediction in Wikipedia. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–

730, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1071.

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. Extracting age-related stereotypes from social media texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3183–3194, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.341.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 2015. doi: 10.48550/ARXIV.1505.07818. URL https://arxiv.org/abs/1505.07818.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.124. URL https://aclanthology.org/2021.findings-acl.124.

S. L. Gordon. *The sociology of sentiments and emotions*, pages 562–569. Rosenberg, M. and Turner, R.H., New York, 1981.

Paul E. Griffiths and Andrea Scarantino. Emotions in the wild: The situated perspective on emotion. In P. Robbins and M. Aydede, editors, *The Cambridge Handbook of Situated Cognition*, chapter 23, pages 437–453. Cambridge University Press, 2005.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. Classifying emotions in customer support dialogues in social media. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 64–73, Los Angeles, September

2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3609. URL https://aclanthology.org/W16-3609.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.11. URL https://aclanthology.org/2020.coling-main.11.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. doi: https://doi.org/10.1111/lnc3.12432. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://aclanthology.org/2020.findings-emnlp.7.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.487. URL https://aclanthology.org/2020.acl-main.487.

Jumayel Islam, Robert E. Mercer, and Lu Xiao. Multi-channel convolutional neural network for Twitter emotion and sentiment recognition. 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 1355–1365,

Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1137. URL https://aclanthology.org/N19-1137.

Carroll E. Izard. The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, 2(4):363–370, 2010. doi: 10.1177/1754073910374661. URL https://doi.org/10.1177/1754073910374661.

William James. II.—WHAT IS AN EMOTION ? *Mind*, os-IX(34):188–205, 04 1884. ISSN 0026-4423. doi: 10.1093/mind/os-IX.34.188. URL https://doi.org/10.1093/mind/os-IX.34.188.

Olga Kellert and Md Mahmud Uz Zaman. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.14. URL https://aclanthology.org/2022.lchange-1.14.

Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508, 2018. URL http://arxiv.org/abs/1805.04508.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 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, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6206. URL https://aclanthology.org/W18-6206.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823.

Richard S. Lazarus. Progress on a Cognitive-Motivational-Relational Theory of Emotion. *The American psychologist*, 46 8:819–34, 1991.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1099.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. 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 4590–4600, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1466. URL https://aclanthology.org/D19-1466.

Kristen A Lindquist. The role of language in emotion: existing evidence and future directions. *Current Opinion in Psychology*, 17:135–139, 2017. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2017.07.006. URL https://www.sciencedirect.com/science/article/pii/S2352250X16302275. Emotion.

Kristen A. Lindquist, Ajay B. Satpute, Tor D. Wager, Jochen Weber, and Lisa Feldman Barrett. The Brain Basis of Positive and Negative Affect: Evidence from a Meta-Analysis of the Human Neuroimaging Literature. *Cerebral Cortex*, 26(5):1910–1922, 01 2015. ISSN 1047-3211. doi: 10.1093/cercor/bhv001. URL https://doi.org/10.1093/cercor/bhv001.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. 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 3469–3483, Online, August 2021. Association for Computational

Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1070.

Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42, 2017. doi: 10.1109/ICDMW.2017.12.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocon. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-srw.26. URL https://aclanthology.org/2021.acl-srw.26.

Saif Mohammad. 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, Portland, OR, USA, jun 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-1514.

Saif Mohammad. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/N12-1071.

Saif Mohammad and Felipe Bravo-Marquez. 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, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5205. URL https://aclanthology.org/W17-5205.

Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-0204.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 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, jun 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2607. URL https://aclanthology.org/W14-2607.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1003. URL https://aclanthology.org/S16-1003.

Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*,

51(4):480–499, 2015. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2014.09.003. URL https://www.sciencedirect.com/science/article/pii/S0306457314000880.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111, 2014. doi: 10.1109/TAFFC.2014.2317187.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Isar Nejadgholi and Svetlana Kiritchenko. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.20. URL https://aclanthology.org/2020.alw-1.20.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 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, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Robert Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*,

21(4-5):529–553, 1982. doi: 10.1177/053901882021004003. URL https://doi.org/10.1177/053901882021004003.

Robert Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344, July 2001. doi: 10.1511/2001.4.344.

Chris Pool and Malvina Nissim. 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, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-4304.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0404. URL https://aclanthology.org/W16-0404.

Changqin Quan and Fuji Ren. An exploration of features for recognizing word emotion. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 922–930, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1104.

Patrick John Ramos, Kiki Ferawati, Kongmeng Liew, Eiji Aramaki, and Shoko Wakamiya. Emotion analysis of writers and readers of Japanese tweets on vaccinations. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 95–103, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wassa-1.10. URL https://aclanthology.org/2022.wassa-1.10.

G. Ramsay. *Analysis and theory of the emotions*. Longman, Brown, Green and Longmans, London, UK, 1848.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.756.

Dante Razo and Sandra Kübler. Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 70–78, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.9. URL https://aclanthology.org/2020.alw-1.9.

Carley Reardon, Sejin Paik, Ge Gao, Meet Parekh, Yanling Zhao, Lei Guo, Margrit Betke, and Derry Tanti Wijaya. BU-NEmo: an affective dataset of gun violence news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2507–2516, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.267.

James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

James A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110 (1):145–72, 2003.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.10. URL https://aclanthology.org/2020.alw-1.10.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

Andrea Scarantino and Ronald de Sousa. Emotion. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021. URL https://plato.stanford.edu/archives/sum2021/entries/emotion/.

Klaus R. Scherer. Appraisal considered as a process of multi-level sequential checking. 92:92–120, 2001.

Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005. doi: 10.1177/0539018405058216. URL https://doi.org/10.1177/0539018405058216.

Klaus R. Scherer and Johnny J. R. Fontaine. 186Driving the emotion process: The Appraisal component. In *Components of Emotional Meaning: A sourcebook*. Oxford University Press, 08 2013. ISBN 9780199592746. doi: 10.1093/acprof:oso/9780199592746.003.0013. URL https://doi.org/10.1093/acprof:oso/9780199592746.003.0013.

Klaus R Scherer and Harald G Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310, 1994.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 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, Copen-

hagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5203. URL https://aclanthology.org/W17-5203.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.468. URL https://aclanthology.org/2020.acl-main.468.

Craig A. Smith and Phoebe C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–38, 1985. doi: https://doi.org/10.1037/0022-3514.48.4.813.

Craig A. Smith and Richard S. Lazarus. *Emotion and Adaptation*, chapter 8, pages 609–637. Guilford, New York, 1990.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1359. URL https://aclanthology.org/P19-1359.

Maximilian Spliethöver and Henning Wachsmuth. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online, December 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.argmining-1.9.

Maximilian Spliethöver and Henning Wachsmuth. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online, December 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.argmining-1.9.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association*

for Computational Linguistics, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://aclanthology.org/P19-1164.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.406.

Carlo Strapparava and Alessandro Valitutti. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Yi Chern Tan and L. Elisa Celis. *Assessing Social and Intersectional Biases in Contextualized Word Representations*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Peggy A. Thoits. The sociology of emotions. *Annual Review of Sociology*, 15:317–342, 1989. ISSN 03600572, 15452115. URL http://www.jstor.org/stable/2083229.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1391. URL https://aclanthology.org/P19-1391.

Enrica Troiano, Laura Oberländer, and Roman Klinger. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1–72, 03 2023. ISSN 0891-2017. doi: 10.1162/coli_a_00461. URL https://doi.org/10.1162/coli_a_00461.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, and Yi Chang. Eliminating sentiment bias for aspect-level sentiment classification with unsupervised opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3002–3012, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.258. URL https://aclanthology.org/2021.findings-emnlp.258.

William Yang Wang, Sameer Singh, and Jiwei Li. Deep adversarial learning for NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 1–5, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5001. URL https://aclanthology.org/N19-5001.

Maximilian Wich, Jan Bauer, and Georg Groh. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.7. URL https://aclanthology.org/2020.alw-1.7.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL https://aclanthology.org/N19-1060.

Anna Wierzbicka. Defining emotion concepts. *Cognitive Science*, 16 (4):539–581, 1992. doi: https://doi.org/10.1207/s15516709cog1604_4.

URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1604_4.

Anna Wierzbicka. The relevance of language to the study of emotions. *Psychological Inquiry*, 6(3):248–252, 1995. ISSN 1047840X, 15327965. URL http://www.jstor.org/stable/1449441.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Xiao Xu, Gert Stulp, Antal Van Den Bosch, and Anne Gauthier. Understanding narratives from demographic survey data: a comparative study with multiple neural topic models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 33–38, Abu Dhabi, UAE, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlpcss-1.4.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1066. URL https://aclanthology.org/N16-1066.

Samira Zad, Joshuan Jimenez, and Mark Finlayson. Hell hath no fury? correcting bias in the NRC emotion lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 102–113, Online, August 2021. Association for Computational

Linguistics. doi: 10.18653/v1/2021.woah-1.11. URL https://aclanthology.org/2021.woah-1.11.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 1–13, Abu Dhabi, UAE, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlpcss-1.1.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. Why do you feel this way? summarizing triggers of emotions in social media posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.642.

Qimin Zhou and Hao Wu. NLP at IEST 2018: BiLSTM-attention and LSTM-attention via soft voting in emotion classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 189–194, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6226. URL https://aclanthology.org/W18-6226.

Yftah Ziser and Roi Reichart. Pivot based language modeling for improved neural domain adaptation. 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 1241–1251, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1112. URL https://aclanthology.org/N18-1112.

A Implementation details

	# topics	# masked topic words
ISEAR	10	100
SSEC	11	110
TALES	10	10
CROWD-ENVENT	8	80
APPREDDIT	10	100
ENISEAR	13	130

Table 15: Number (#) of topics and the resulting number of removed (i.e., masked) topic words.

B Topic Modeling

	Emotion						
Topic	Anger	Disgust	Fear	Joy	Sadness	Shame	Σ_{Topic}
Outlier	414	423	254	281	248	792	2412
Love	228	108	28	318	226	303	1211
Exams	59	37	129	362	98	146	831
Death	20	30	71	73	398	49	641
Shame	127	47	14	8	30	399	625
School	74	55	20	23	17	209	398
Animals	21	146	83	7	50	23	330
Alcohol	26	192	24	1	4	79	326
Accidents	51	28	148	5	8	85	325
Fear	4	3	293	5	9	9	323
Theft	72	27	31	11	8	95	244
Σ_{Emotion}	1096	1096	1095	1094	1096	2189	7666

Table 16: The number of instances in ISEAR per topic×emotion combination with the overall instance count for each topic in the last column (Σ_{Topic}) and for each emotion in the last row (Σ_{Emotion}).

	BERTopic		LDA
ID	Label	Top 4	Top 4
0	*outlier*	hillaryclinton, hillary, barackobama, obama	help, said, follow, equality
1	feminism	feminism, feminist, feminists, feminismiscruelty	president, people, abortion, think
2	prayer	rosary, prayer, prayers, mary	feminists, world, love, like
3	abortion	abortion, abortions, fetus, unborn	bless, don, white, like
4	climate	climate, warming, mission, environment	going, obama, unborn, live
5	clinton	hillaryclinton, hilary, hillary, hillaryforsc	hillaryclinton, make, right, rt
6	twitter	tweet, twitter, excited, hashtags	choice, rt, country, hillaryclinton
7	trump	realdonaldtrump, donaldtrump, trump, donald	love, pay, like, just
8	gay marriage	gaymarriage, marriageequality, scotusmarriage, ...	like, feminist, potus, donald
9	latino	hispanic, latino, racistublicans, latinos	scotus, good, right, don
10	swearing	yell, dontjudgeme, fuck, bitches	bad, think, need, life
11	patriotism	liberty, patriot, patriotswillrise, tyranny	

Table 17: Each topics (ID) most informative keywords (Top 4) found in SSEC by the BERTopic and LDA topic models with ID 0 denoting the BERTopic outlier class. The manual label for each topic (Label) only applies to BERTopic because of the erratic nature of the LDA topics.

	BERTopic		LDA
ID	Label	Top 4	Top 4
0	*outlier*	flowers, daughter, garden, water	great, took, ran, red
1	birds	bird, sang, birds, singing	good, princess, thought, bed
2	flowers	flowers, blossoms, flower, blossom	king, think, time, yes
3	tabitha twitchit	duchess, ribby, tabitha, kitten	away, queen, time, went
4	old english	thou, thee, thy, mercy	cat, little, flew, turned
5	piggies	pigling, pig, piggies, pigs	took, opened, house, little
6	royalty	princesses, princess, queen, princes	mother, great, flowers, voice
7	dressmaking	tailor, dressed, garments, shop	long, trees, green, wood
8	hansel & gretel	hans, hansel, carpenter, shepherd	saw, street, house, came
9	boat	boat, sail, sailed, sailing	said, day, people, bread
10	predators	fox, foxy, wolf, lion	left, bird, seen, pocket
11			fell, asleep, night, happened
12			tod, said, little, room

Table 18: Each topics (ID) most informative keywords (Top 4) found in Tales by the BERTopic and LDA topic models with ID 0 denoting the BERTopic outlier class. The manual label for each topic (Label) only applies to BERTopic because of the erratic nature of the LDA topics.

	# Topics	∅ Topic	StD	Example topics	Outlier
ISEAR	10	525	290	love, exams, death, theft	2412
SSEC	11	305	219	feminism, prayer, climate, hillary clinton	1513
Tales	10	388	183	old pronouns, flowers, princess, hansel&gretel	6457
crowd-enVENT	8	584	298	promotion, covid, relationships, accidents	1925
APPReddit	10	43	12	depression, relationships, suicide, love	352
enISEAR	13	58	25	accidents, eating, holidays, work	245

Table 19: Number (#), average size (∅), standard deviation (StD), and examples of the topics as found by BERTopic for all corpora. All numbers exclude the outlier class, whose size is given last.

	BERTopic		LDA
ID	Label	Top 4	Top 4
0	*outlier*	won, game, play, win	ex, time, feel, help
1	feelings	feel, feeling, feelings, past	went, graduated, won, game
2	promotion	promotion, promoted, accomplishment, workplace	really, didn, test, told
3	relationships	married, friendship, fiance, relationship	best, new, told, away
4	covid	covid, coronavirus, vaccination, positive	good, years, did, wasn
5	dogs	dog, dogs, puppy, barking	day, just, boss, got
6	graduation	graduated, degree, university, graduating	friend, holiday, cancer, asked
7	pregnancy	birth, born, pregnant, pregnancy	didn, took, day, mum
8	driving	accident, driving, drove, car	having, friend, didn, really
9			secret, car, able, son
10			offered, money, birth, son
11			dad, didn, went, party

Table 20: Each topics (ID) most informative keywords (Top 4) found in CROWD-ENVENT by the BERTopic and LDA topic models with ID 0 denoting the BERTopic outlier class. The manual label for each topic (Label) only applies to BERTopic because of the erratic nature of the LDA topics.

	BERTopic		LDA
ID	Label	Top 4	Top 4
0	*outlier*	friend, hope, friends, feels	today, family, hope, year
1	depression	depression, depressed, anxiety, therapist	years, getting, night, completely
2	everyday life	laundry, job, jobs, morning	just, little, ve, actually
3	driving	restaurant, driving, radio, taco	doing, life, ve, time
4	love	love, loving, loves, loved	make, need, ve, like
5	romantic relationships	breakup, gf, girlfriend, relationship	years, control, fucking, year
6	reddit	reddit, hate, posts, downvoted	comments, fuck, people, really
7	anger	anger, angry, rage, frustrated	feel, time, needs, day
8	death	suicide, suicidal, overdose, died	ve, maybe, love, make
9	platonic relationships	friends, classmates, friendzone, socialized	thought, friend, love, think
10	vaccination	autism, vaccines, parents, children	life, fucking, today, want
11			does, things, think, just
12			ve, need, fucking, time
13			today, thank, love, edit

Table 21: Each topics (ID) most informative keywords (Top 4) found in APPReddit by the BERTopic and LDA topic models with ID 0 denoting the BERTopic outlier class. The manual label for each topic (Label) only applies to BERTopic because of the erratic nature of the LDA topics.

	BERTopic		LDA
ID	Label	Top 4	Top 4
0	*outlier*	dog, homeless, past, walked	job, bought, took, daughter
1	death	pain, died, heart, funeral	trying, friend, going, night
2	dogs	barking, dog, pet, dogs	son, phone, husband, got
3	accidents	accident, hurt, driving, drove	couldn, knew, died, dad
4	theft	stole, stealing, shopping, inadvertently	drove, year, day, hit
5	birth	feeling, birth, newborn, grandchild	day, dog, people, wrong
6	food	eating, tasted, smelled, sausage	drink, mother, friend, day
7	affairs	cheated, affair, cheating, married	home, passed, like, school
8	UK politics	brexit, protesters, laughing, refugees	documentary, said, friend, watched
9	christmas	christmas, xmas, feelings, holidays	thought, home, friends, time
10	bullying	bullied, bullying, kids, kid	job, lost, children, having
11	work	workplace, job, employment, fired	tv, door, lost, started
12	relationships	friends, unkind, betrayed, meet	walked, man, dog, died
13	spooky	darkness, dark, haunted, night	good, help, decision, school
14			driving, ate, saw, old
15			took, buy, did, birthday
16			went, home, day, friend
17			feel, ate, friend, going
18			children, family, just, daughter

Table 22: Each topics (ID) most informative keywords (Top 4) found in enISEAR by the BERTopic and LDA topic models with ID 0 denoting the BERTopic outlier class. The manual label for each topic (Label) only applies to BERTopic because of the erratic nature of the LDA topics.

C Corellations between Topics and Emotion/Appraisal Annotations

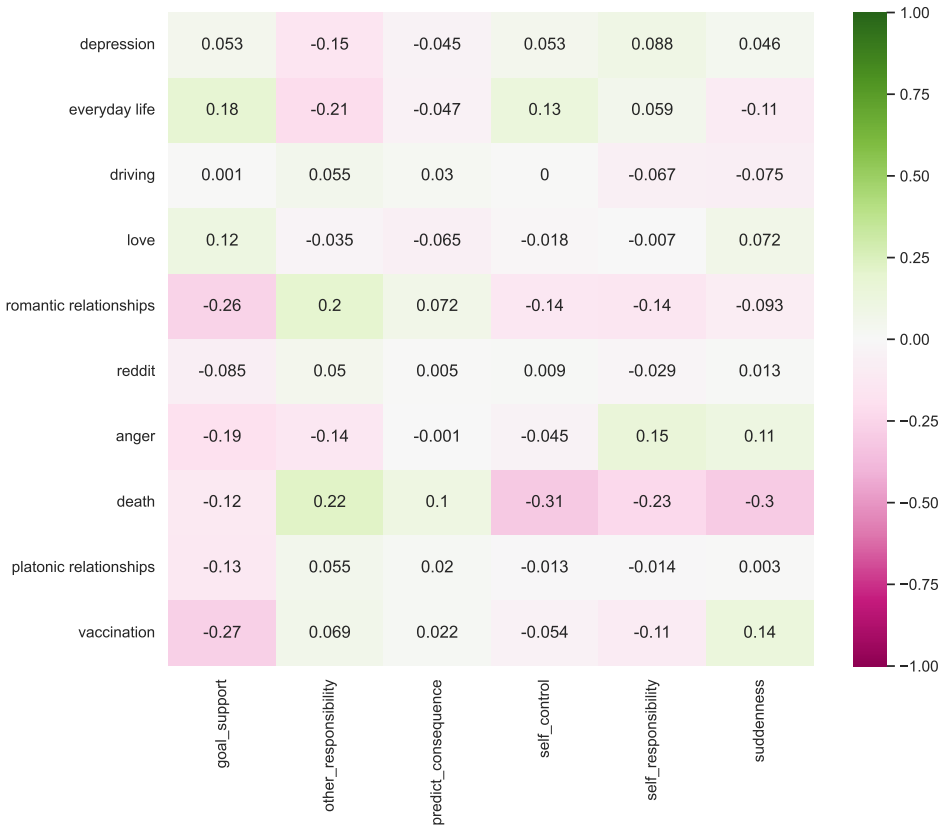

Figure 15: Normalized pointwise mutual information between topics and appraisal annotations in APPREDDIT.

Figure 16: Normalized pointwise mutual information between topics and appraisal annotations in ENISEAR.

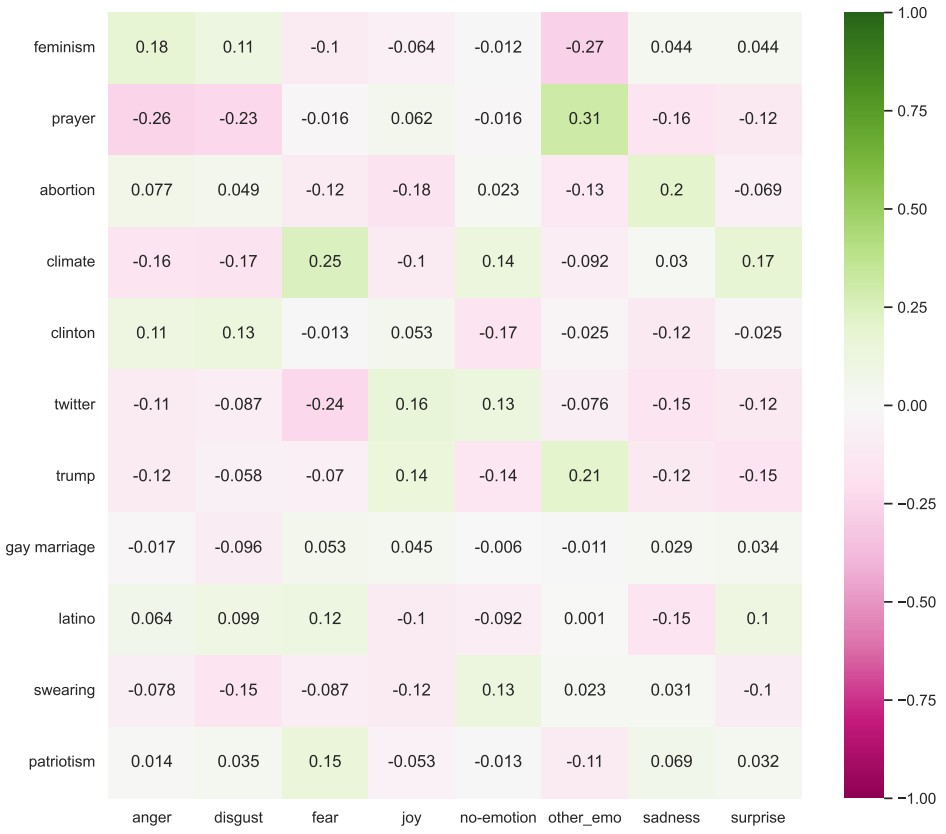

Figure 17: Normalized pointwise mutual information between topics and emotion annotations in SSEC.

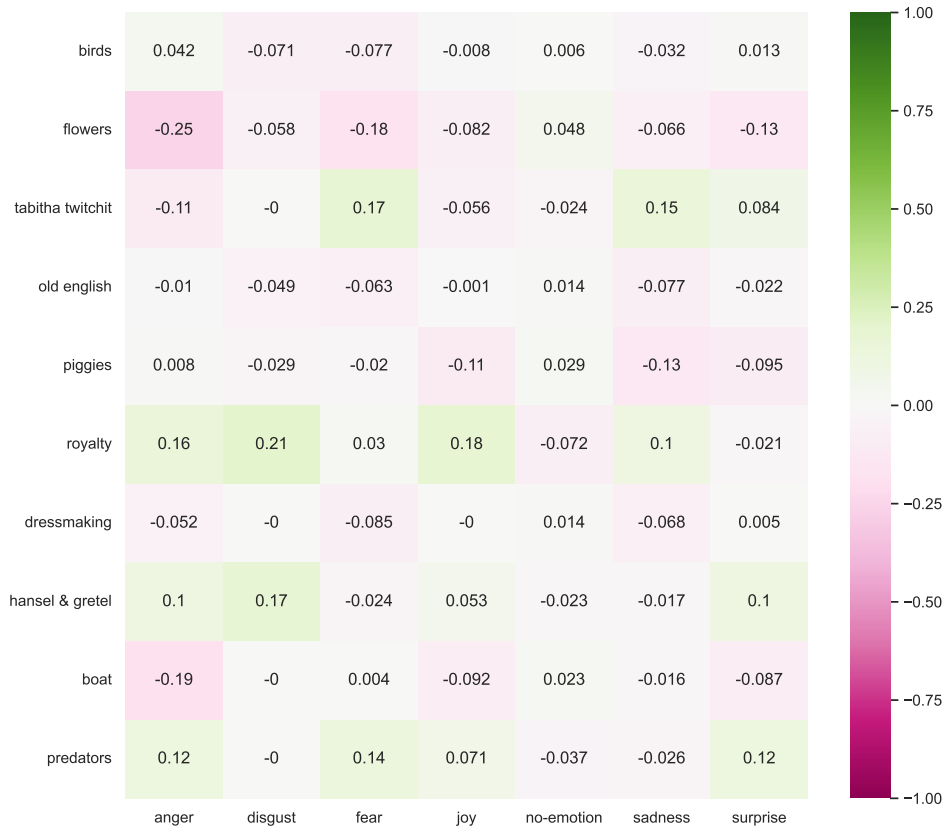

Figure 18: Normalized pointwise mutual information between topics and emotion annotations in TALES.

D Results

D.1 CrossTopic/InTopic

D.2 CrossTopic-mask/InTopic-mask

D.3 CrossTopic-grl/InTopic-grl

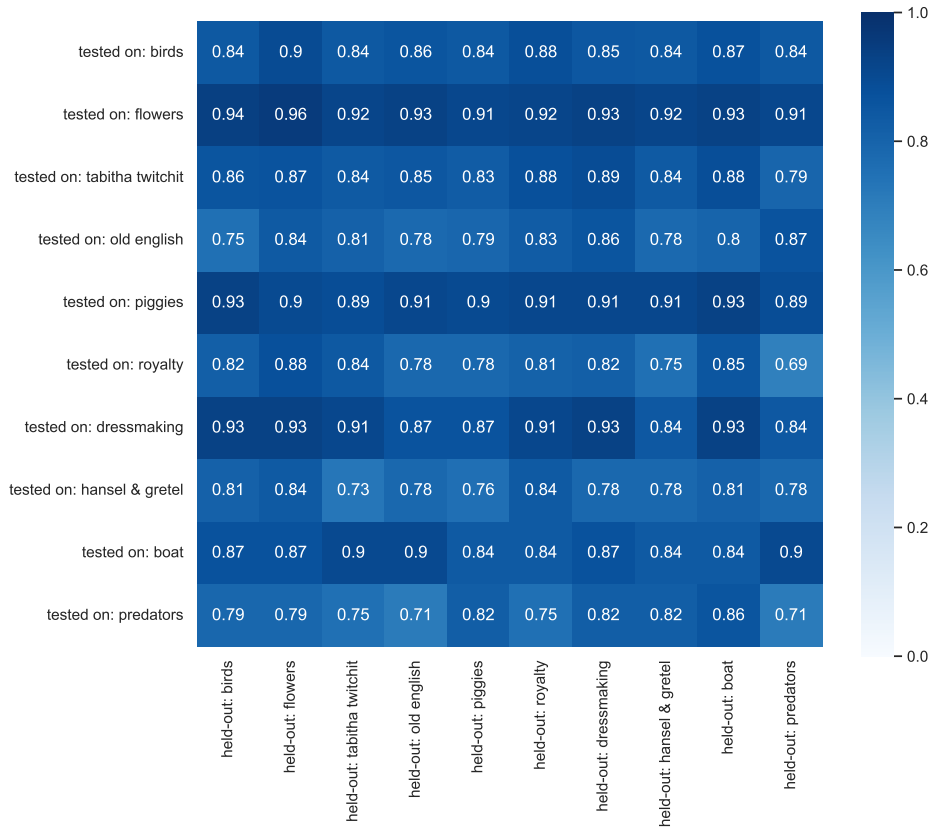

Figure 19: Micro-average F_1 for each testset in TALES, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

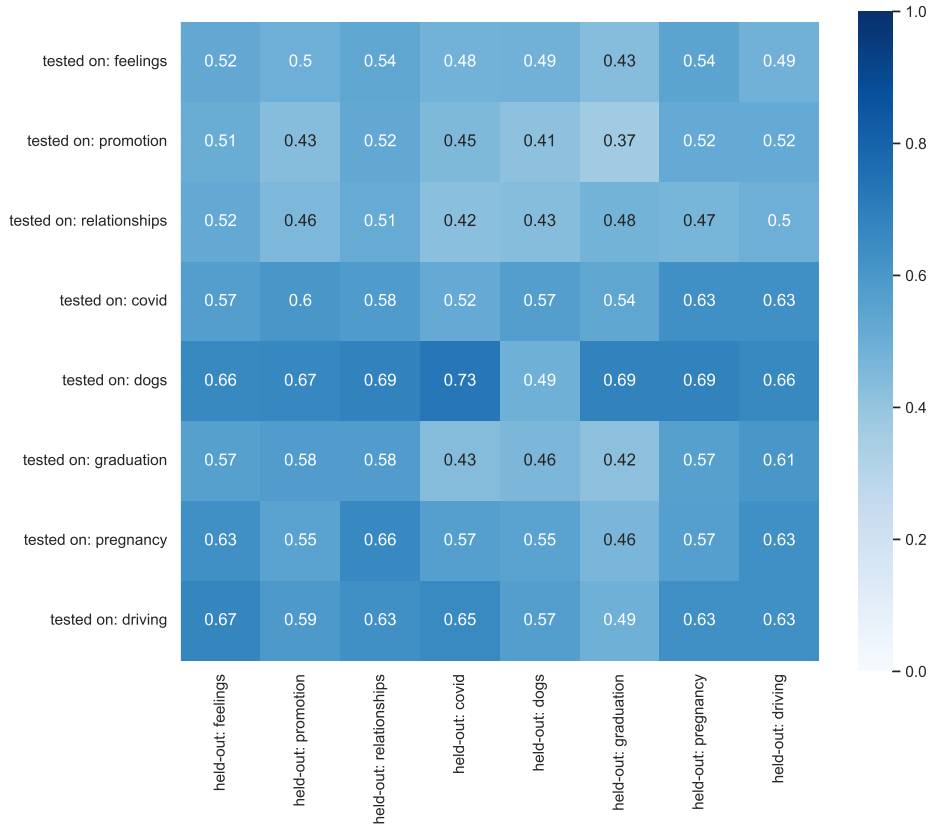

Figure 20: Micro-average F_1 for each testset in CROWD-ENVENT (emo), in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

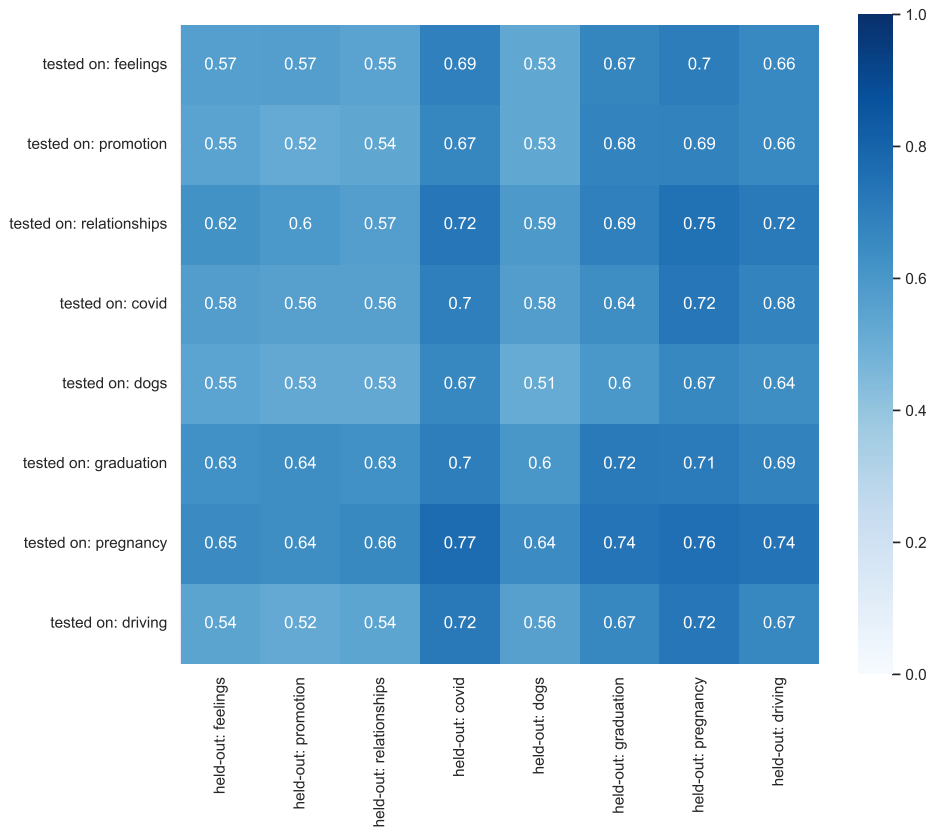

Figure 21: Micro-average F_1 for each testset in CROWD-ENVENT (appraisal), in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

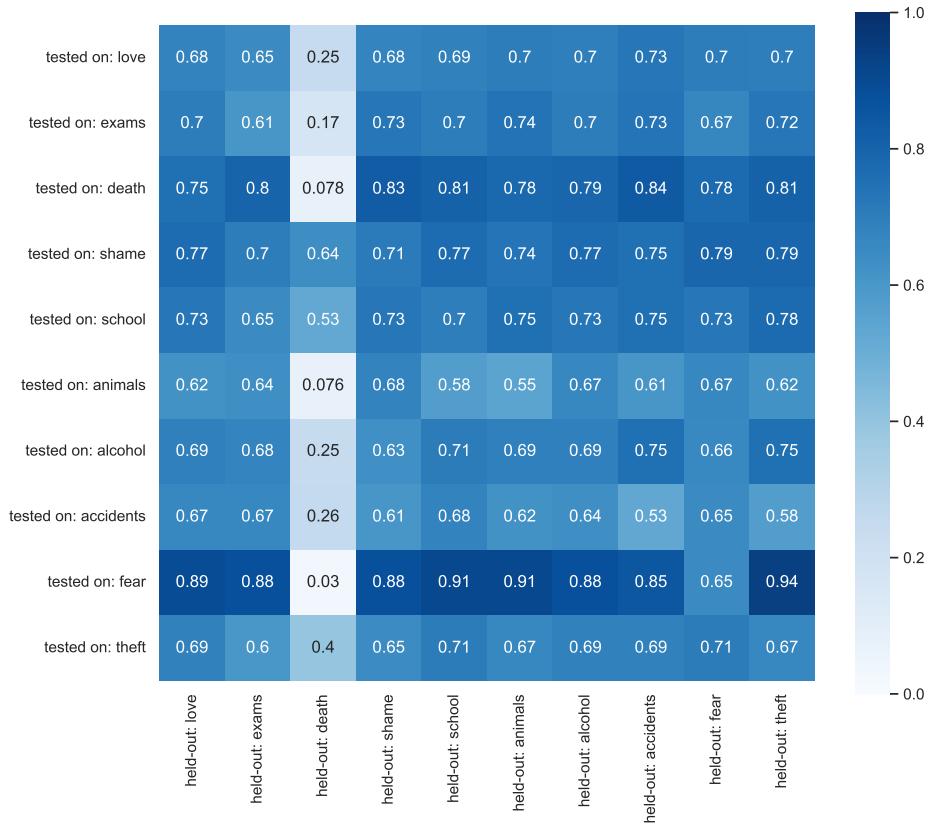

Figure 22: Micro-average F_1 for each testset in ISEAR, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

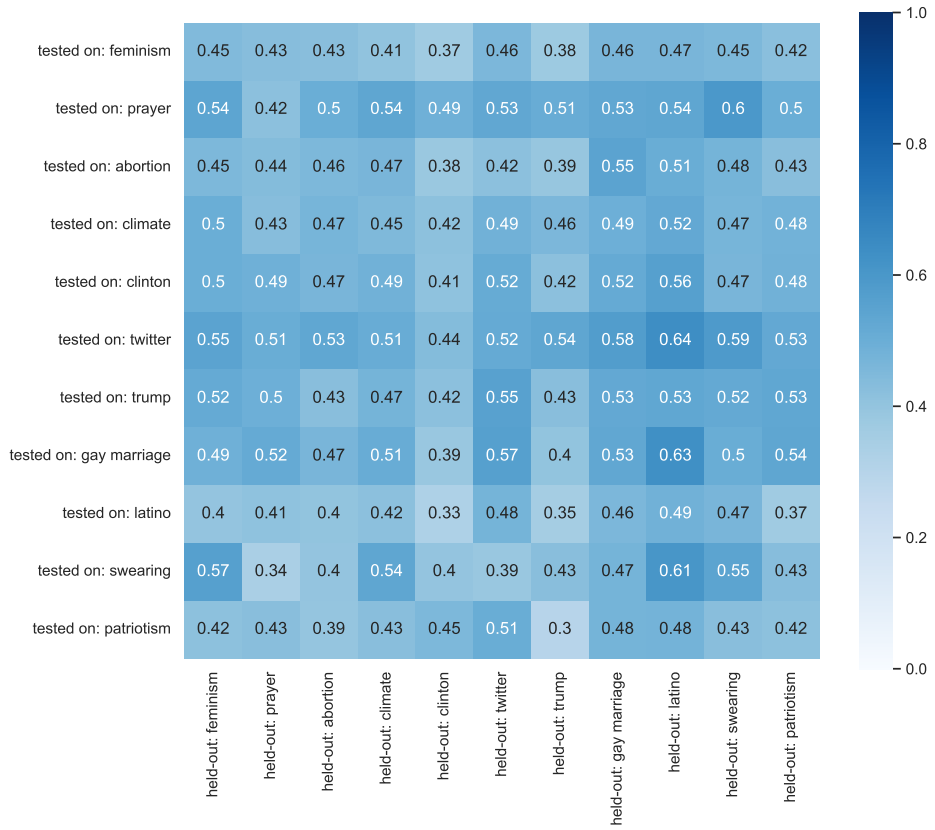

Figure 23: Micro-average F_1 for each testset in SSEC, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

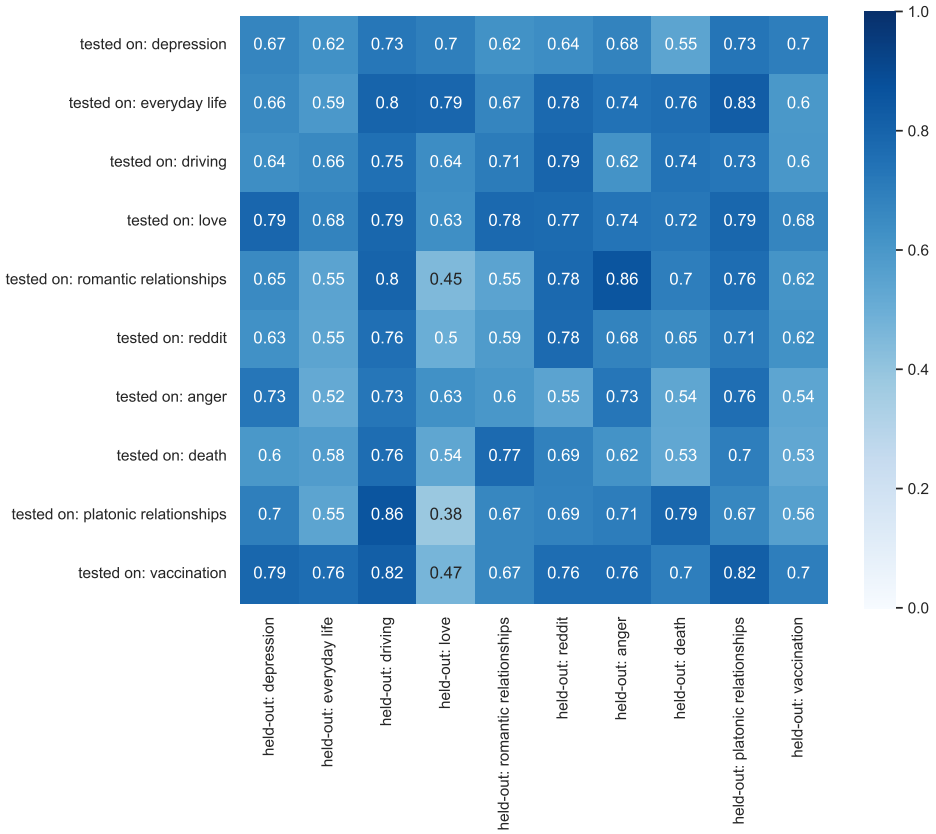

Figure 24: Micro-average F_1 for each testset in APPREDDIT, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

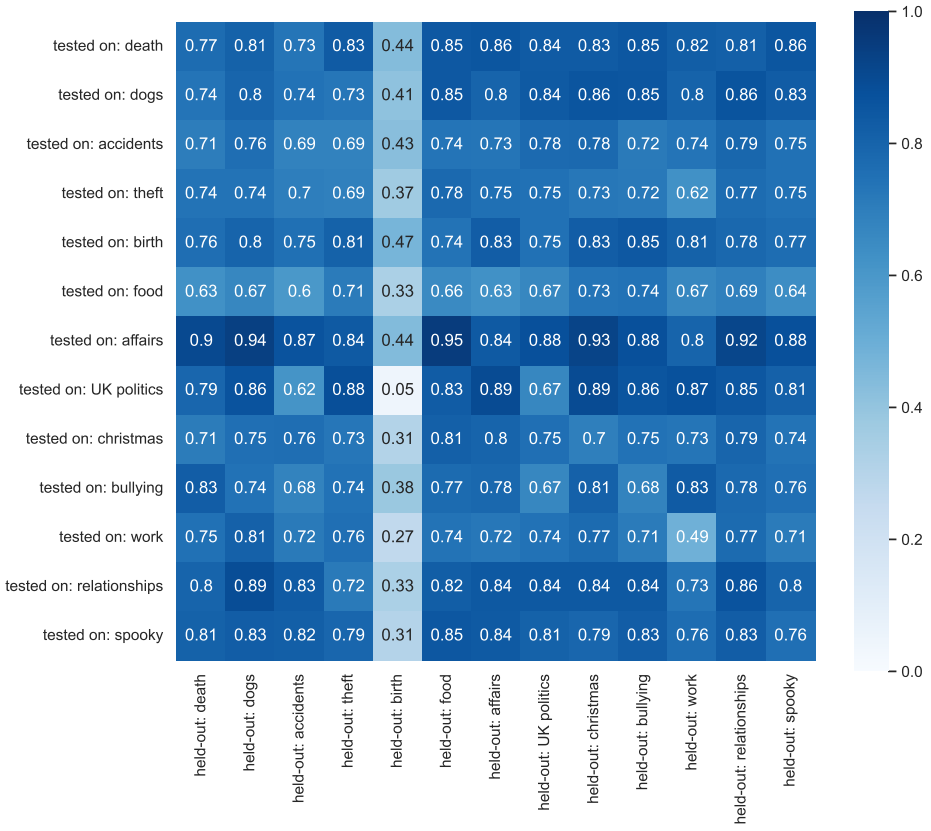

Figure 25: Micro-average F_1 for each testset in ENISEAR, in relation to held-out topic (CROSSTOPIC/INTOPIC setting).

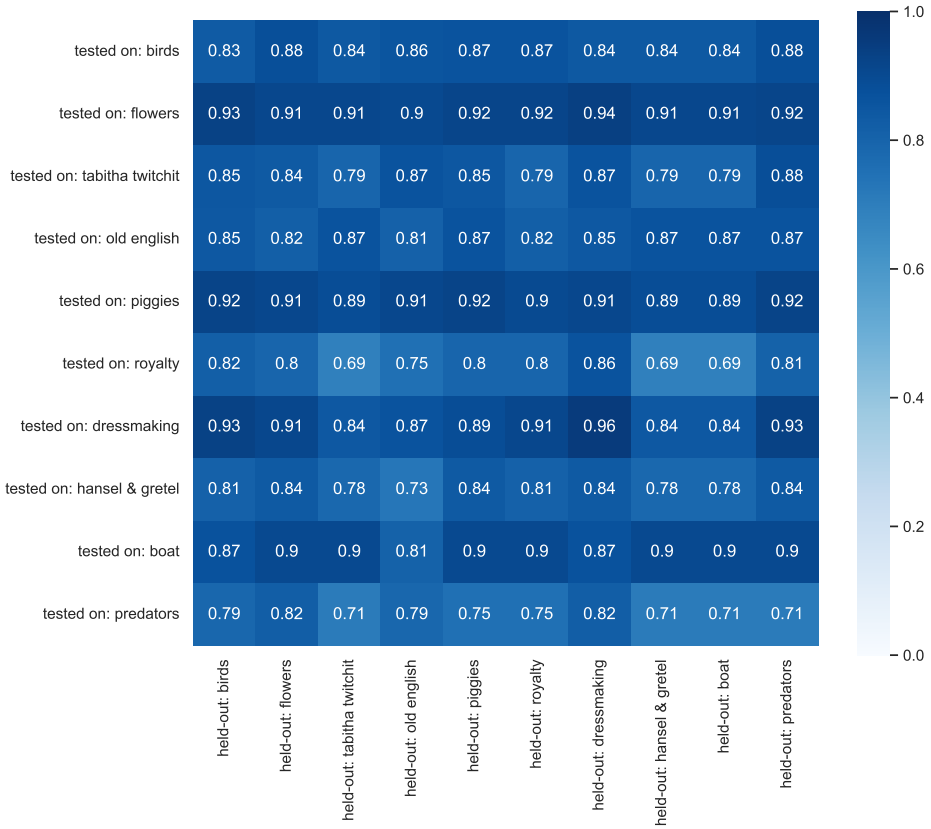

Figure 26: Micro-average F_1 for each testset in TALES, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

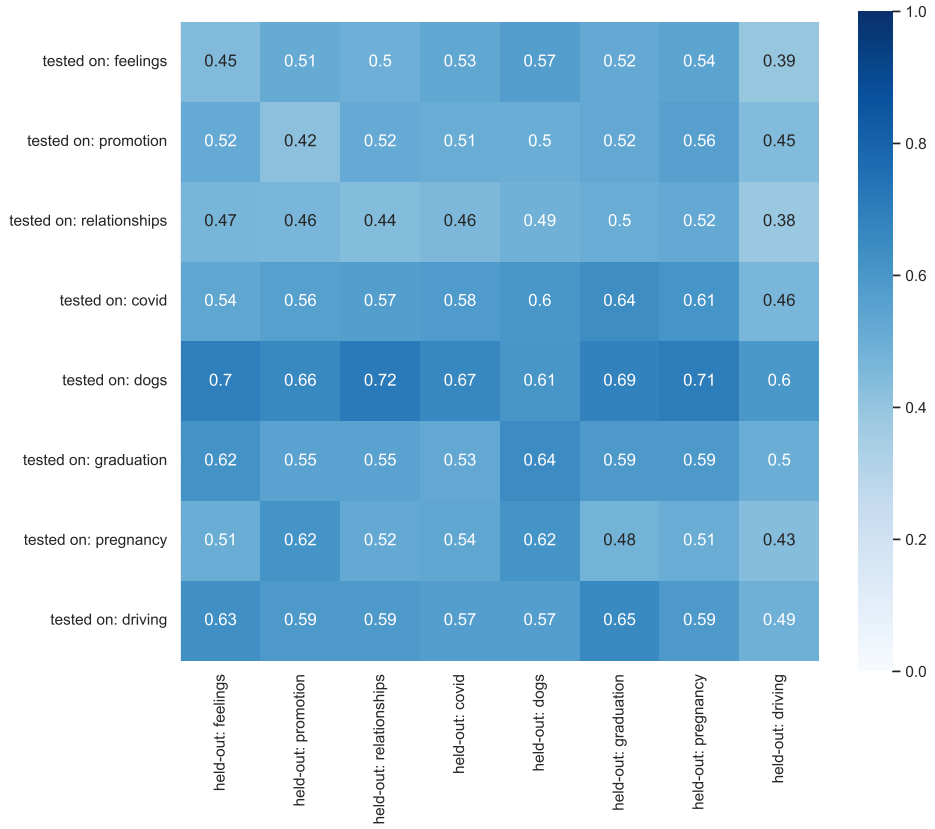

Figure 27: Micro-average F_1 for each testset in CROWD-ENVENT (emo), in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

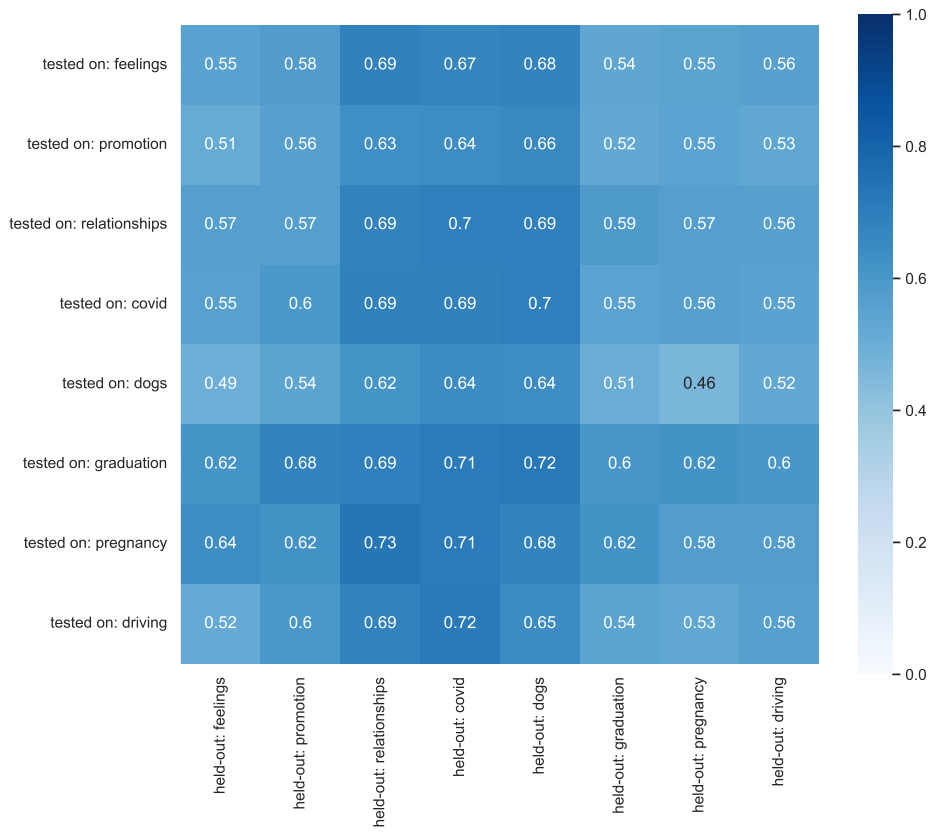

Figure 28: Micro-average F_1 for each testset in CROWD-ENVENT (appraisal), in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

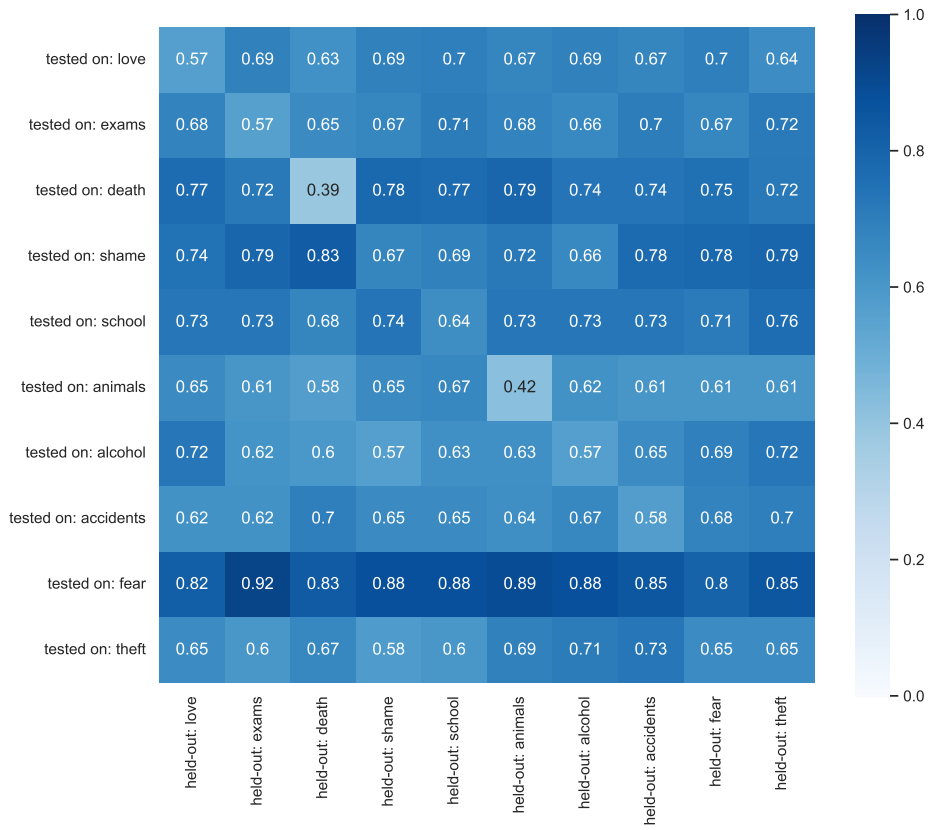

Figure 29: Micro-average F_1 for each testset in ISEAR, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

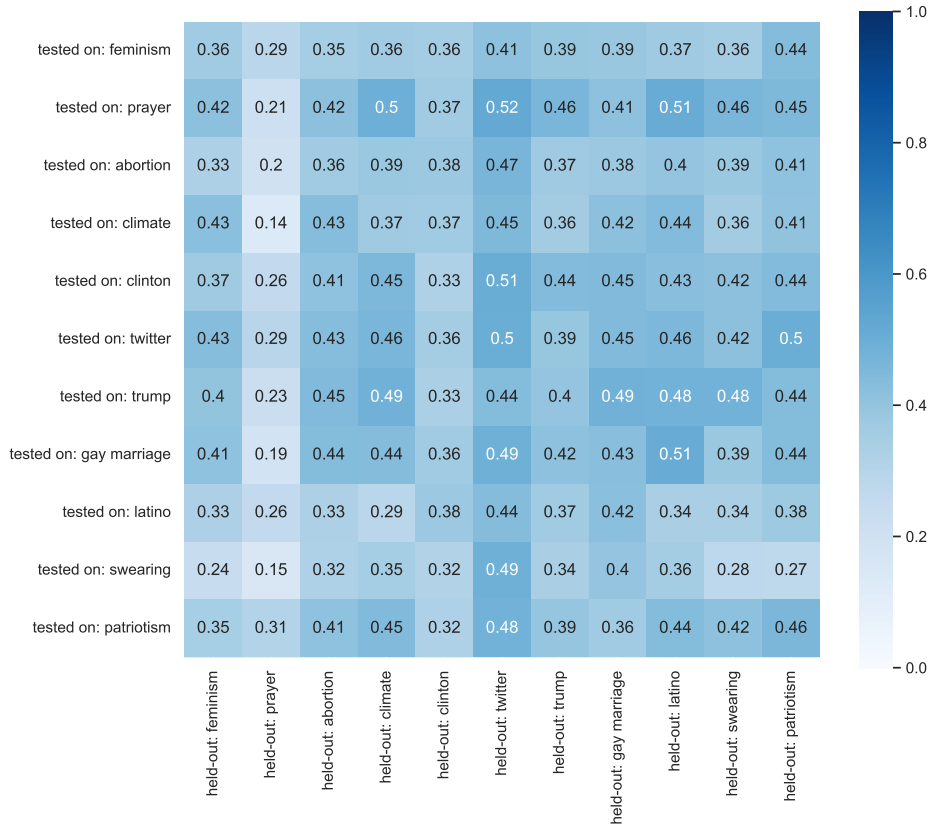

Figure 30: Micro-average F_1 for each testset in SSEC, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

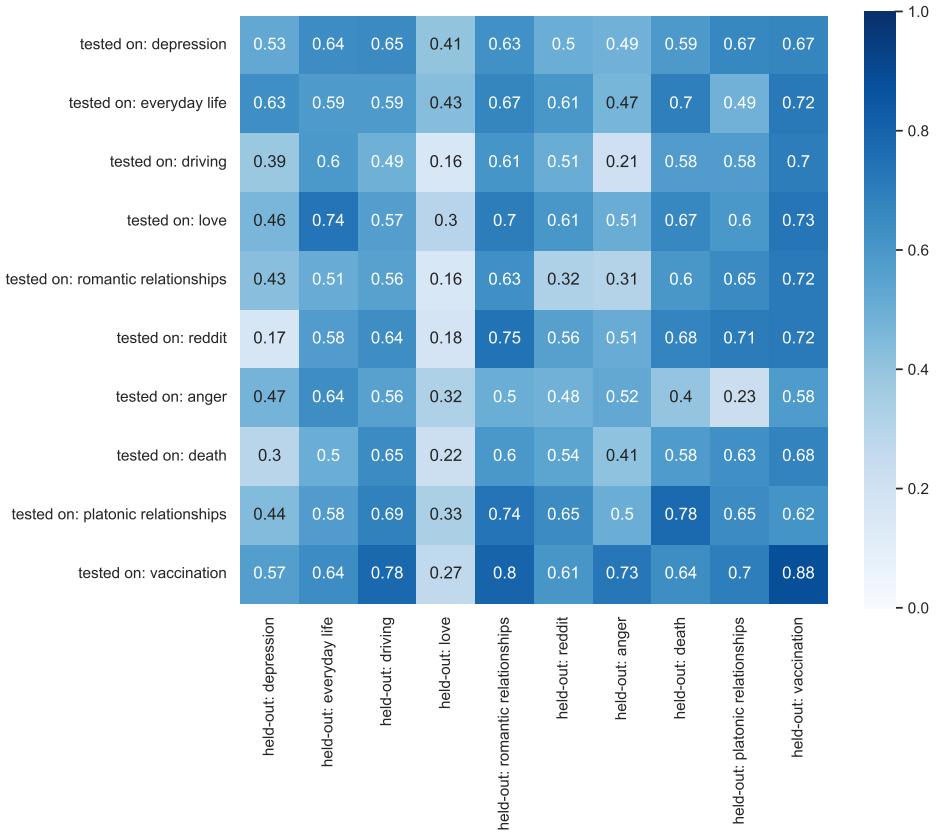

Figure 31: Micro-average F_1 for each testset in APPREDDIT, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

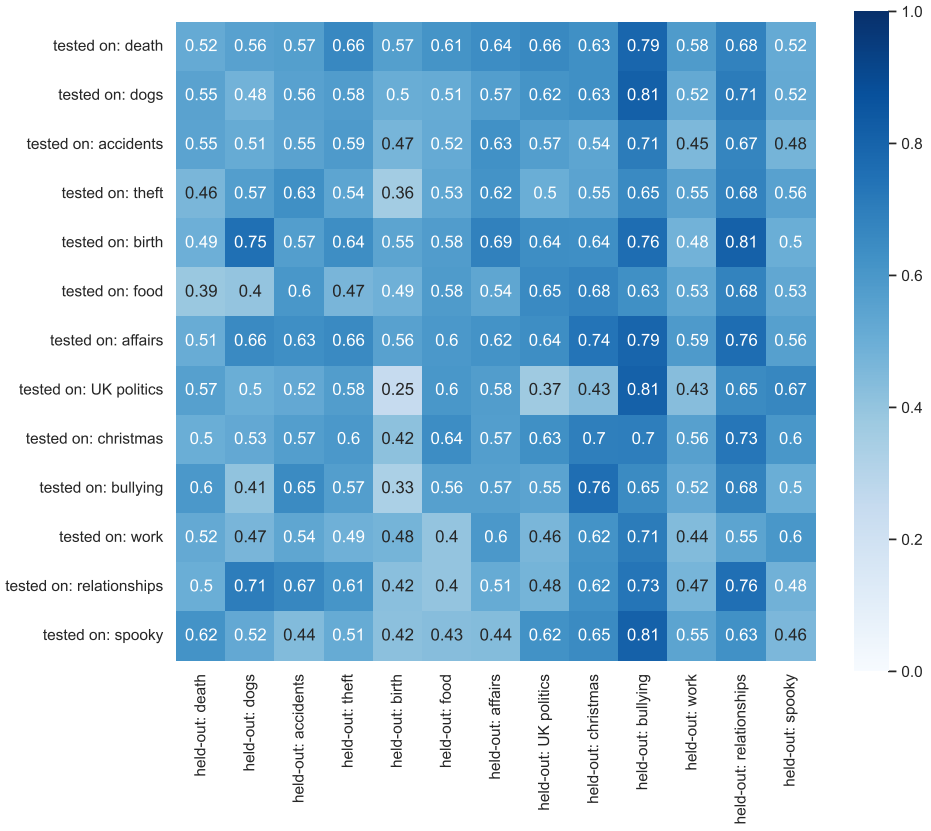

Figure 32: Micro-average F_1 for each testset in ENISEAR, in relation to held-out topic (CROSSTOPIC-MASK/INTOPIC-MASK setting).

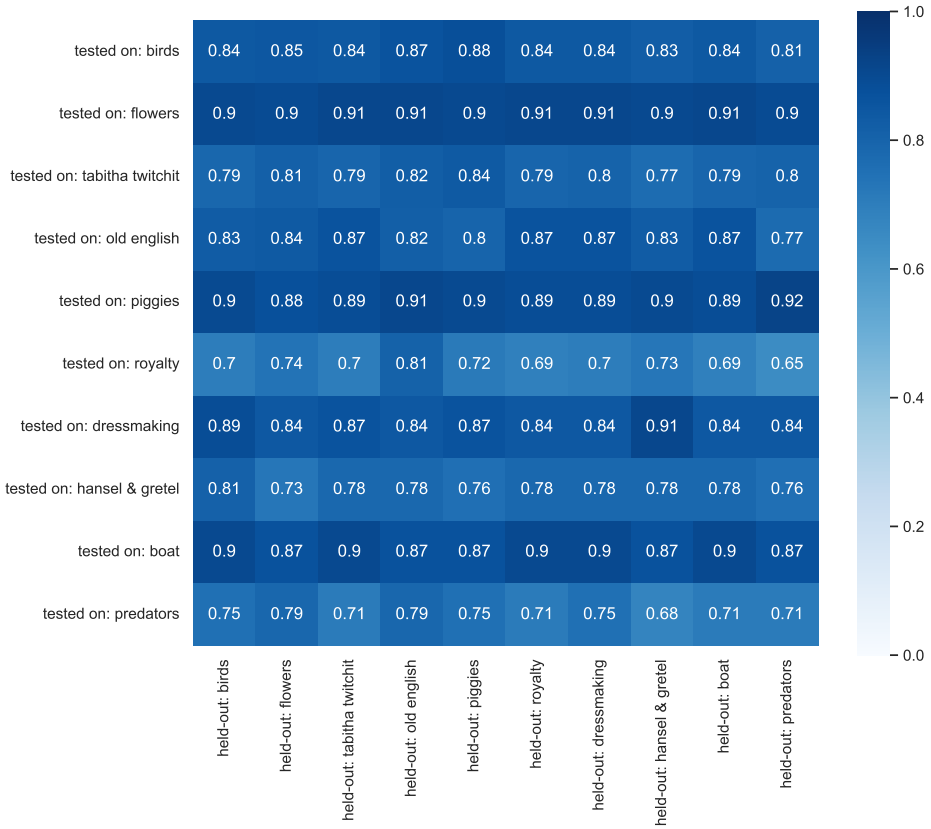

Figure 33: Micro-average F_1 for each testset in TALES, in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

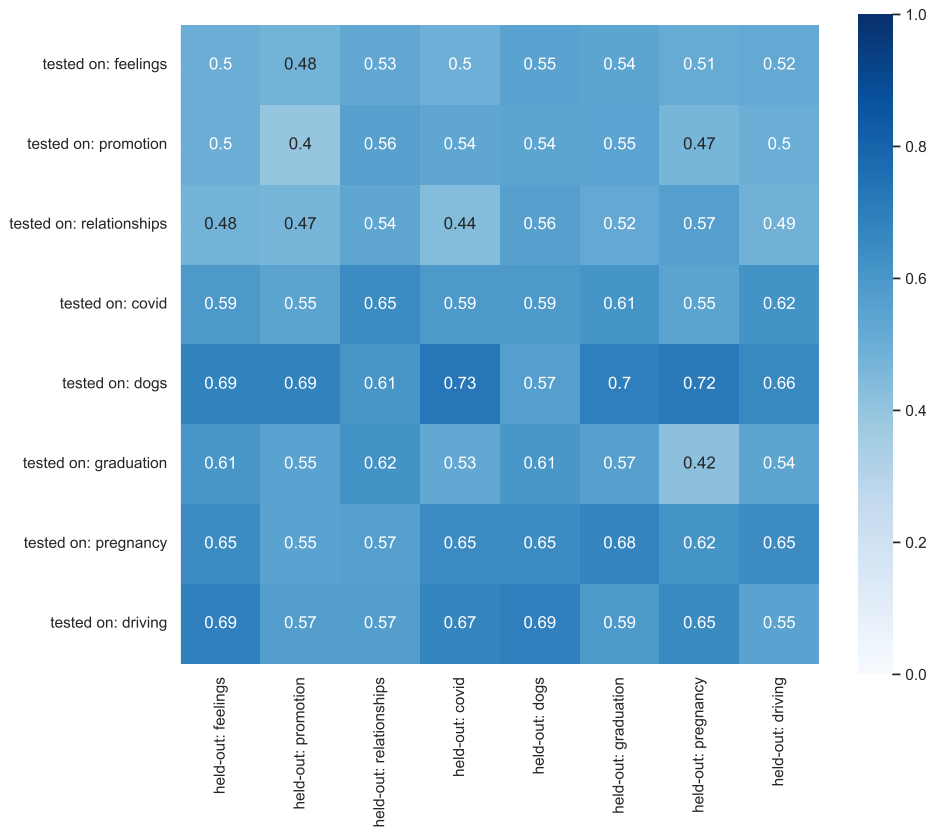

Figure 34: Micro-average F_1 for each testset in CROWD-ENVENT (emo), in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

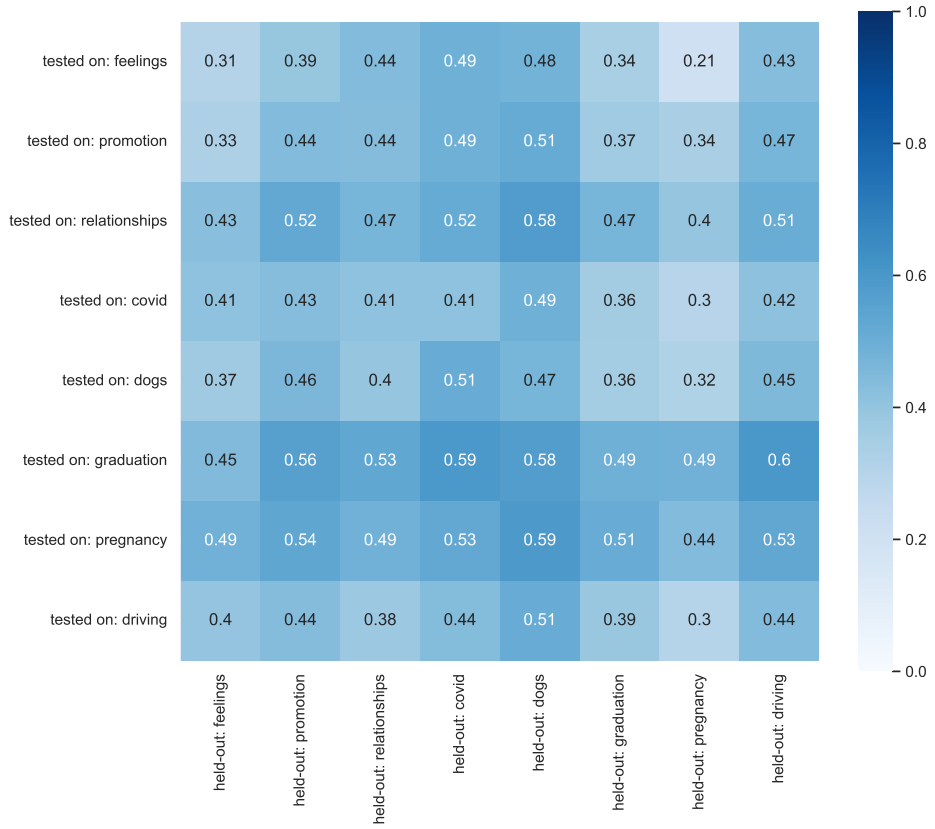

Figure 35: Micro-average F_1 for each testset in CROWD-ENVENT (appraisal), in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

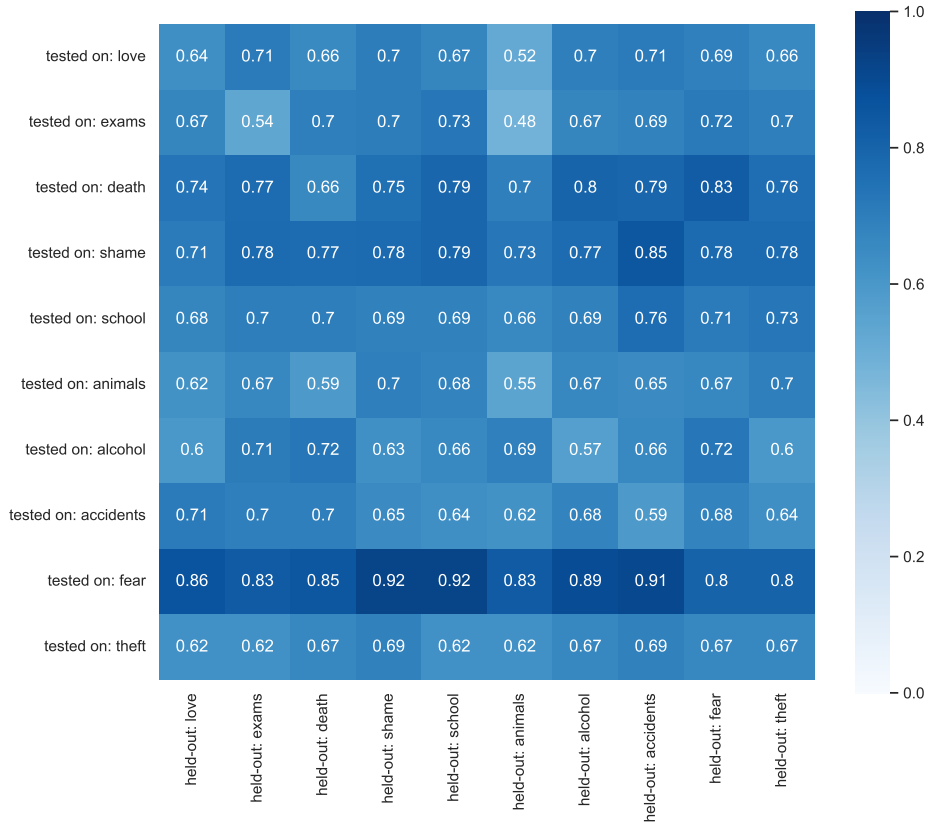

Figure 36: Micro-average F_1 for each testset in ISEAR, in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

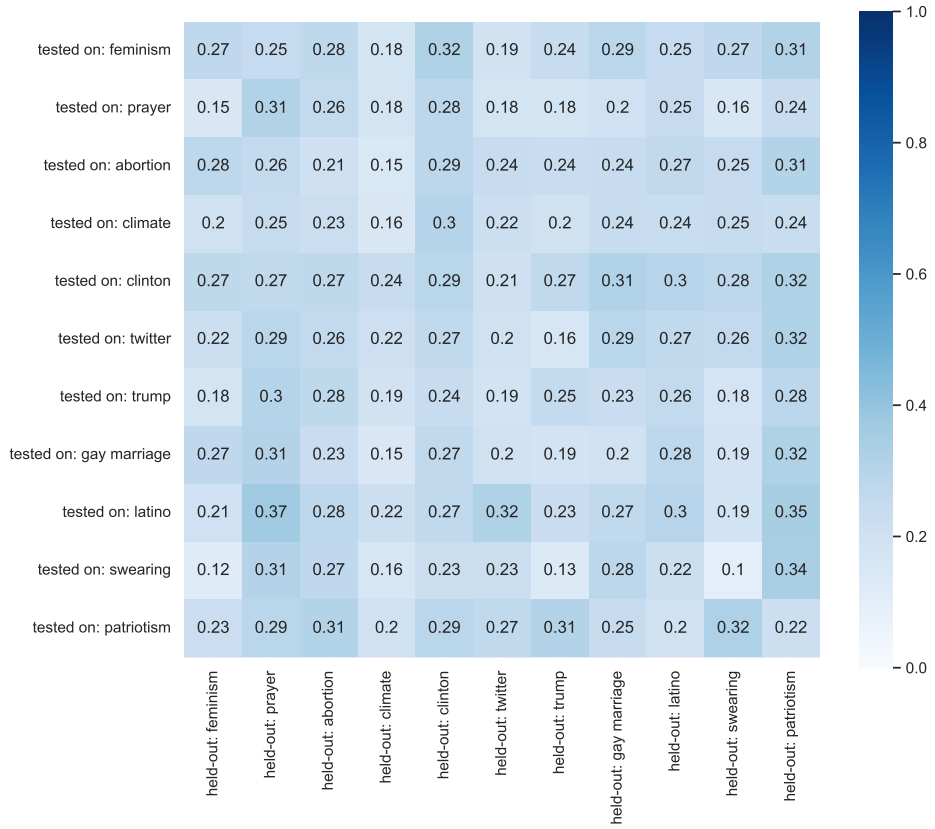

Figure 37: Micro-average F_1 for each testset in SSEC, in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

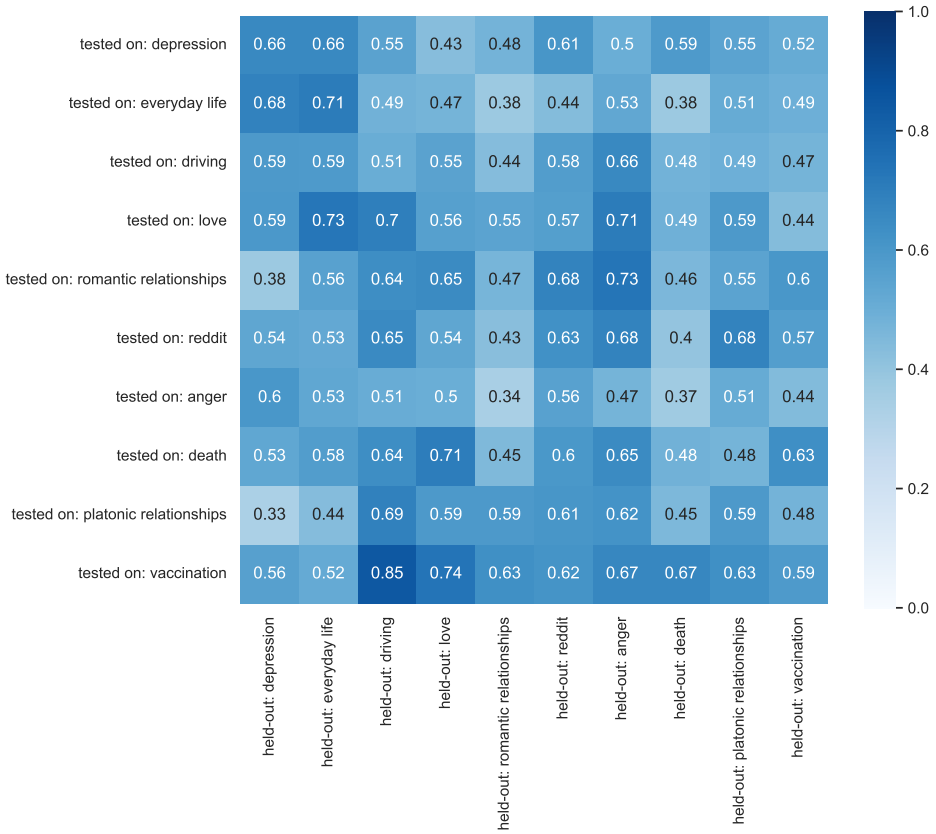

Figure 38: Micro-average F_1 for each testset in APPREDDIT, in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).

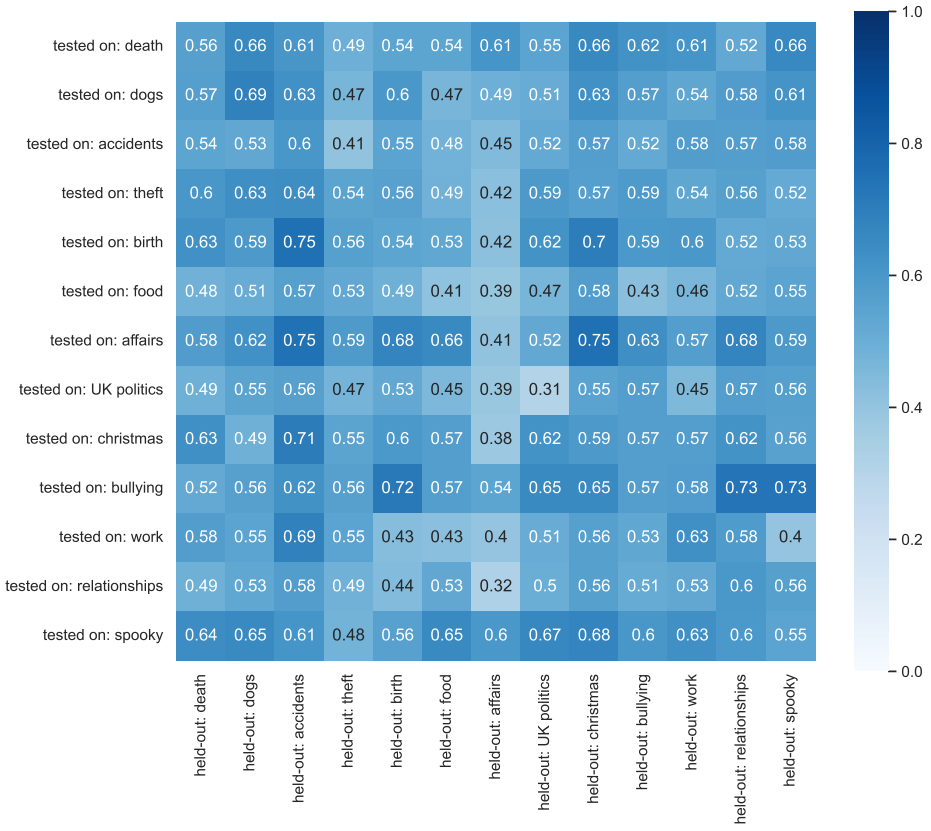

Figure 39: Micro-average F_1 for each testset in ENISEAR, in relation to held-out topic (CROSSTOPIC-GRL/INTOPIC-GRL setting).