Emotions in Literature: Computational Modelling in the Context of Genres and Characters

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

Vorgelegt von
Evgeny Kim
aus Chorvoq Taschkent, Usbekistan

Hauptberichter	Prof. Dr. Sebastian Padó
Mitberichter	Prof. Dr. Heike Zinsmeister

Tag der mündlichen Prüfung: 24. 08. 2020
Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2020

Erklärung (Statement of Authorship)
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig 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.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my endless gratitude to Roman Klinger, my first and foremost supervisor, for being my mentor and with whom we spent long hours discussing research, and whose insightful guidelines helped me to stay on track and retain the research-oriented agenda in my work.

I am thankful to Sebastian Padó, who always found the time to listen to the story of my thesis from start to end bringing valuable suggestions on research directions and never failing to raise exciting questions and engaging in mindful discussions.

I would like to thank Nils Reiter and Sarah Schulz for helping me navigating in an intricate world of digital humanities in the first year of my PhD program.

I thank all of my peers and colleagues with whom I co-wrote papers and shared office space. To name a few, I am thankful to Laura Ana Maria Bostan, Enrica Troiano, Camilo Thorne, Abhijeet Gupta and the rest members of the TCL group. Without you it wouldn't be possible for me to keep my head above water during all those years.

List of Abbreviations

biLSTM ..... Bidirectional long-short term memory

BoW ..... Bag-of-words

CNN ..... Convolutional neural network

COREF ..... Coreference

CRF ..... Conditional random fields

DH ..... Digital humanities

GRU ..... Gated recurrent unit

IAA ..... Inter-annotator agreement

IOB ..... Inside-outside-beginning

LSTM ..... Long-short term memory

MaxEnt ..... Maximum entropy

MEntity ..... Masked entity

MLP ..... Multilayer perceptron

MRole ..... Masked role

NE ..... Named entity

NER ..... Named entity recognition

NLP ..... Natural language processing

NoInd. ..... No indicator

PMI ..... Point-wise mutual information

RF ..... Random forest

SVM ..... Support vector machines

VAD ..... Valence-arousal-dominance

List of Publications

Parts of the research described in this thesis have been published in:

- **Chapter 2**: Kim, E. and Klinger, R. (2019c). A survey on sentiment and emotion analysis for computational literary studies. *ZfDG*, 4

- **Chapter 3**: Kim, E., Padó, S., and Klinger, R. (2017a). Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics

- **Chapter 4**: Kim, E. and Klinger, R. (2018). Who feels what and why? Annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics

- **Chapter 5**: Kim, E. and Klinger, R. (2019b). Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics

Other related publications that are not described in this work:

- Kim, E., Padó, S., and Klinger, R. (2017b). Prototypical Emotion Developments in Literary Genres. In *Digital Humanities 2017: Conference Abstracts*, pages 288–291, Montréal, Canada. McGill University and Université de Montréal

- Köper, M., Kim, E., and Klinger, R. (2017). IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark. Association for Computational Linguistics

- Barth, F., Kim, E., Murr, S., and Klinger, R. (2018). A reporting tool for relational visualization and analysis of character mentions in literature. In *Book of Abstracts – Digital Humanities im deutschsprachigen Raum*, Cologne, Germany

- Kim, E. and Klinger, R. (2019a). An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on*

Storytelling, pages 56–64, Florence, Italy. Association for Computational Linguistics

- Haider, T., Eger, S., Kim, E., Klinger, R., and Menninghaus, W. (2020). PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resources Association (ELRA)

- Bostan, L. A. M., Kim, E., and Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resources Association (ELRA)

Abstract (English)

A significant body of research on the role of emotions in the literature indicates that emotions can be an integral part of a literary text. Recently, the study of emotions has entered the research agenda of computational literary studies giving rise to a new type of studies based both on qualitative and quantitative insights. This thesis builds on top of previous work, contributing to a better understanding of certain aspects of emotions in a literary text. The first half of the thesis focuses on the textual properties of emotions and their link to literary genres. The author investigates whether emotion dictionaries can be used for emotion classification of a literary text and if "emotional arcs" derived with dictionaries are useful features in genre classification. The results of the genre classification experiments suggest that both emotional arc–and emotion content words in general–are a good predictor of the genre though worse than bag-of-words. The author then examines the structure of emotion, mainly, which parts of the structure are the most difficult from the annotation and classification viewpoint. As the results of the experiments suggest, emotion *causes* are

the most difficult parts of the structure, which is explained by the fact that they are mostly events. At the same time, such parts of the structure as *emotion* and *experiencer* are the least difficult to model, especially in joint learning setting. In the second part of this work, the author shifts the focus to emotion classification of interpersonal character relationships. Specifically, he investigates which model architectures are optimal for relationship classification and finds that recurrent neural architecture with positional indicators suits the task the best. Finally, the last experimental chapter shows that real-world application of the model to an arbitrary text is feasible, but requires a parameter optimization process, which does not necessarily lead to high precision and recall. On the whole, the work outlines several prospects in the computational study of emotion in literature and enables both literary scholars and NLP engineers to make use of this thesis' insights in their work.

Abstract (Deutsch)

Eine bedeutende Anzahl von Untersuchungen zur Rolle von Emotionen in der Literatur legt nahe, dass Emotionen ein wesentlicher Bestandteil eines literarischen Textes sein können. In jüngster Zeit ist das Studium der Emotionen in die Forschungsagenda der rechnergestützten Literaturwissenschaft eingegangen, was zu einer neuen Art von Studien geführt hat, die sowohl auf qualitativen als auch auf quantitativen Erkenntnissen beruhen. Diese Arbeit baut auf früheren Arbeiten auf und trägt zu einem besseren Verständnis bestimmter Aspekte bei von Emotionen in einem literarischen Text. Die erste Hälfte der Arbeit befasst sich mit den textlichen Eigenschaften von Emotionen und ihrer Verbindung zu literarischen Genres. Der Autor untersucht, ob Emotionswörterbücher zur Emotionsklassifizierung eines literarischen Textes verwendet werden können und ob mit Wörterbüchern abgeleitete "emotionale Bögen" nützliche Merkmale bei der Genreklassifizierung sind. Die Ergebnisse der Genreklassifizierungsexperimente legen nahe, dass sowohl der emotionale Bogen - als auch die Wörter mit Emotionsinhalt im Allgemeinen ein guter Prädiktor für das

Genre sind, wenn auch schlechter als eine Wortsammlung. Der Autor untersucht dann hauptsächlich die Struktur der Emotionen, welche Teile der Struktur unter dem Gesichtspunkt der Annotation und Klassifizierung am schwierigsten sind. Wie die Ergebnisse der Experimente zeigen, sind Emotionen *Ursachen* die schwierigsten Teile der Struktur, was durch die Tatsache erklärt wird, dass es sich hauptsächlich um Ereignisse handelt. Gleichzeitig sind Teile der Struktur wie *Emotion* und *Erfahrene* am wenigsten schwer zu modellieren, insbesondere beim gemeinsamen Lernen. Im zweiten Teil dieser Arbeit verlagert der Autor den Fokus auf die Emotionsklassifizierung zwischenmenschlicher Charakterbeziehungen. Insbesondere untersucht er, welche Modellarchitekturen für die Beziehungsklassifizierung optimal sind, und stellt fest, dass eine rekurrente neuronale Architektur mit Positionsindikatoren am besten zur Aufgabe passt. Schließlich zeigt das letzte experimentelle Kapitel, dass eine reale Anwendung des Modells auf einen beliebigen Text möglich ist, jedoch einen Parameteroptimierungsprozess erfordert, der nicht unbedingt zu hoher Präzision und Rückruf führt. Insgesamt skizziert die Arbeit mehrere Perspektiven für die rechnergestützte Untersuchung von Emotionen in der Literatur und ermöglicht es sowohl Literaturwissenschaftlern als auch NLP-Ingenieuren, die Erkenntnisse dieser These in ihrer Arbeit zu nutzen.

1 Motivation and Overview

1.1 Introduction and Motivation

Much of our daily experience influences and is influenced by the emotions we feel (Schwarz, 2000). This experience is not limited to real events. People can feel emotions because they are reading a novel or watching a play or a movie (Johnson-Laird and Oatley, 2016; Djikic et al., 2009). There is a growing body of literature that pinpoints the importance of emotions for literary comprehension (Robinson, 2005; Hogan, 2010; Bal and Veltkamp, 2013; Djikic et al., 2013; Samur et al., 2018; Van Horn, 1997), as well as research that recognizes the deliberate choices people make with regard to their emotion states when seeking narrative enjoyment, for example a book or a film (Zillmann et al., 1980; Bryant and Zillmann, 1984; Ross, 1999; Oliver, 1993; Mar et al., 2011). Discussions about the role of emotions in literature find its place not only in academic writings but in creative writing tutorials as well. Such books as *Writing fiction for dummies* (Ingermanson and Economy, 2009) and *The emotion thesaurus* (Ackerman and Puglisi,

2012) teach beginner writers to build believable emotion descriptions and evoke emotion response in a reader.

The study of emotions is not confined to English and Psychology university departments. *Affective computing* is a field of study embracing recognition, modelling, and generation of emotions by a computer (Picard, 2000). It is a vibrant area of research and engineering spanning a broad range of applications, from emotion recognition from face to voice, body postures and gestures, images and text (Calvo and D'Mello, 2010). There are application programming interfaces (APIs) for emotion recognition developed by big companies such as Microsoft[1] and IBM[2]. Even Amazon, the biggest online retailer, is working on a technology that can read human emotions (Day, 2019). Despite some criticism (Barrett et al., 2019) regarding the viability of emotion recognition using a computer, emotion recognition market is still growing and, according to the recent Emotion Detection and Recognition market report[3], the growth is expected to continue.

Literary studies have not escaped the trend of interest in emotion research. Although, the topic of emotions in litera-

[1] https://azure.microsoft.com/en-us/services/cognitive-services/face/

[2] https://www.ibm.com/support/knowledgecenter/en/SSWTQQ_1.2.0/com.ibm.swg.ba.cognos.sifs_solution_guide.1.2.0.doc/c_trd_emotion_detection_library.html

[3] https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market

ture has been addressed in different time periods (*e.g.*, Tolstoy (1962); Baker (1927); Fish (1970); Neill (1991); Yanal (1999); Sætre et al. (2014); Longo (2020)), some believe that an "emotion turn" in literary studies occurred in the late 2000s (Crozier-De Rosa, 2010).

In contrast to other fields that study subjective phenomena[4] computationally, most of the research on emotions in literature was and is done via *close reading* (Hinchman and Moore, 2013), a "reading to uncover layers of meaning that lead to deep comprehension" (Boyles, 2012, p. 90). Close reading has been the main tool of literary analysis since the beginning of the twentieth century and still remains a fundamental method of literary criticism (Jänicke et al., 2015).

An opposite to close reading is the so-called *distant reading* (Moretti, 2005), a term closely related to *digital literary studies* (Hoover et al., 2014), both of which refer to the practice of running a textual analysis on a computer to yield quantitative results. Another closely related term is *computational literary studies* (CLS), which Da (2019) defines as "the statistical representation of patterns discovered in text mining fitted to currently existing knowledge about literature, literary history, and textual production."

Although the concepts of distant reading and CLS are relatively new and owe their existence to a recent "computational

[4]Phenomena based on personal feelings or opinions

turn" within humanities (Vanhoutte, 2013; Jockers and Underwood, 2016), first works that explore emotions in literature using computational methods appeared long before the so-called "digital revolution" (Lanham, 1989).

For example, the first results of a computer-assisted modelling of emotions in literature are presented by Anderson and McMaster (1982). Challenged by the question why some texts are more interesting than others, in their paper, Anderson and McMaster concluded that the "emotion tone" of a story can be responsible for the reader's interest. The results of their study suggest that a large-scale analysis of "emotion tone" of the collection of texts is possible with the help of a computer program.

There are two implications of this finding. First, Anderson and McMaster suggested that by identifying emotion tones of text passages one can model affective patterns of a given text or a collection of texts, which in turn can be used to challenge or test existing literary theories. Second, their approach to *affect* modelling demonstrated that the stylistic properties of texts can be defined on the basis of their emotion interest and not only linguistic characteristics. With regard to these implications, the paper by Anderson and McMaster is an important early research piece as it laid out a "roadmap" for some of the basic applications of sentiment and emotion analysis of texts, namely sentiment and emotion pattern recognition from text

and computational text characterization based on sentiment and emotion.

In my thesis, I will show that computational emotion analysis in literary context provide an interesting angle of analysis and could help literary scholars in verifying hypotheses, as well as language engineers for a variety of tasks. I will explore how emotions are expressed in texts and how challenging their detection by a computer is. We shall see the difficulties of annotating emotions, as well as the richness and diversity of the language of emotion. At the same time, I will present suitable model architectures for the task of emotion detection in literary text. My work combines elements of both distant reading and automatic emotion analysis. In some sense, this is an interdisciplinary study the outcomes of which can be of use for both DH practitioners and natural language engineers.

1.2 A Big Picture: Research Questions

Anderson and McMaster showed possible research directions that could emerge from their study. In my thesis, I continue the tradition of text analysis laid down by Anderson and McMaster.

That being said, I will not confine myself to exploring the ramifications of their claims. As it often happens in research, the answers to one question lead to new questions. In the case

of my thesis, there are at least three big interrelated research directions and associated questions that I will explore. These phenomena are:

1) an "emotion tone" of a literary text in the context of genre;

2) the linguistic structure of emotion in text;

3) emotion relationships of fictional characters.

Each of the above-mentioned directions has one or more associated research questions and tasks, which I present in the subsequent sections. After that, I will conclude the section with a summary of how they fit together forming a single storyline.

1.2.1 Emotion in the Context of Genre

Anderson and McMaster showed that almost any text can be converted to a time-series graph, where a horizontal axis represents chronological story line, and a vertical axis represents the amount of emotion in each time point. However, they never reported any sort of a large-scale analysis that would generalize observations from focused analyses of a few books they performed. As we shall see in Section 2, there are studies that scale Anderson and McMaster work, but they either focus on rather small sets of emotions (*e.g.*, Reagan et al. (2016)'s focus is on *happiness* alone) or compare limited literary genres (*e.g.*, fairytales vs novels as in Mohammad (2011)).

In my thesis, I will explore the notion of an emotion as a

property of a literary text in the context of genre. To be more specific, I will deal with a chronological sequence of emotion in text, coined as the "emotional arcs" (Reagan et al., 2016, p .2). This research direction continues the line of the research dealing with literary genres and other literary text categorizations (Gao et al., 2016; Samothrakis and Fasli, 2015; Reagan et al., 2016; Mohammad, 2011). The general idea behind these works is that different types of literary texts have different emotion arcs and these *emotion arcs* form distinctive patterns.

An emotion arc can be thought of as of a *story* or a *narrative arc*, that is a chronological construction of a plot. The most famous narrative arc was introduced by Freytag (1863) in the form of a pyramid and comprises exposition, rising action, climax, falling action, and denouement (see top of Fig. 1.1). Equally to a story arc, an emotion arc (see bottom of Fig. 1.1) features chronological peaks and throughs that correspond to a degree to which certain emotion is expressed in text. For example, a horror story may start with little or no fear, which can grow reaching a climax and then subsequently subside and revert to normal. But how can we know that a story evolves in this or another way? This is not a perfunctory question, as the idea of analyzing a text though the prism of an emotional arc may be of interest to literary scholars and engineers. For example, a literary scholar working with a diachronic literary corpus could use this analysis to study the trends in the

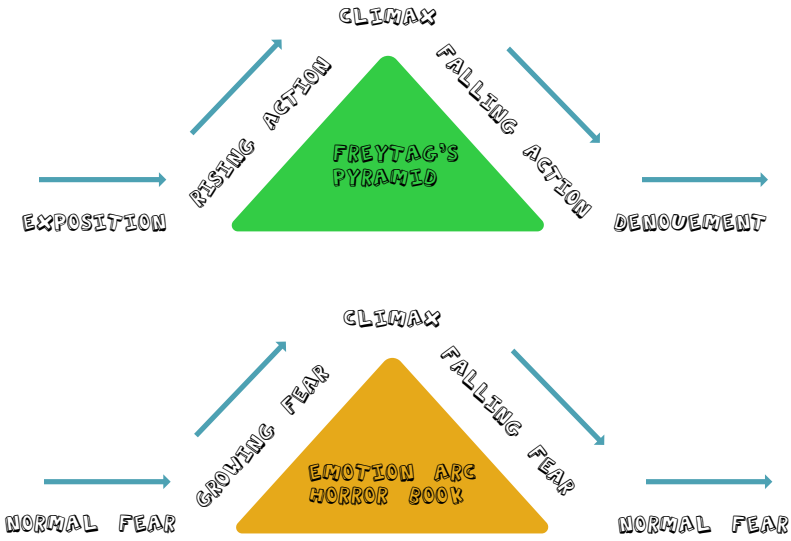


Figure 1.1: Freytag's pyramid of dramatic action (top) and a hypothetical emotion arc showing chronological change in the level of fear in a horror story (bottom).

use of emotion within different literary periods; or a natural language engineer from Amazon may find that such analysis helps in generating high-level emotion profiles of Kindle books automatically.

In my thesis, I aim to validate the gist of Anderson and McMaster's method. Namely, I propose to make sure that an emotion dictionary used to construct an emotion arc is indeed capable of providing meaningful results on a literary text. To that end, I will use a modern *emotion dictionary* to derive emotion scores for different parts of texts. I will then use these scores to construct emotional arcs that will represent each text and help me to answer the following question:

Research Question 1

Does emotion classification performed with dictionaries allow for the classification of genres?

To answer this question, I will need to work up a framework that enables to use dictionaries in a classification task. In this work, I propose to *use an emotion dictionary for emotion arcs construction*. Then, I will use the resulting emotion arcs in a downstream classification task where their effect can be estimated and compared to other analytical tools.

In my thesis, I opt for a *genre* classification scenario, an established research field in computational linguistics. The task of genre classification is formulated as follows: Given a collection of texts and a set of genres the texts belong to, associate each text in the collection to one or more genres from the set. One may wonder how can an emotional arc be used in genre classification. Consider Figure 1.2, which shows a hypothetical chronological change in the *joy* emotion in the genres of *romance*, *adventure*, and *science fiction*. It can be seen that there are some differences – some are barely noticeable, some are pronounced – in the way the emotion is represented in the text. So if there are differences observed with a naked eye, it could mean that the classifier will be able to discern them to make a decision on the genre of the book in question. Again, this is a hypothetical example. As we shall see in Section 3, the differences between genres are not always as clearly distinguishable as in Figure 1.2.

Answering the Research Question 1 would already be sufficient to understand whether emotion analysis with dictionaries is a useful tool of text analysis. However, the question of why they are useful will still remain unanswered. What parts of the text exactly are responsible for the emotion expression? What is the linguistic structure behind an emotion in text? Are there any patterns in how emotions are expressed in text? And, finally, can we find these patterns and *analyze emotions*

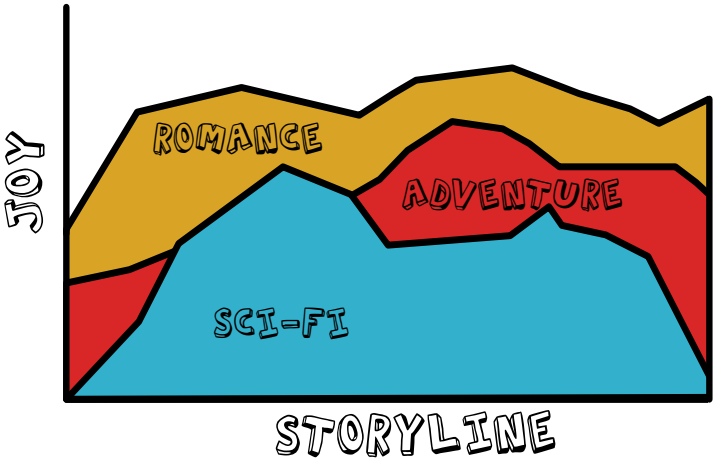


Figure 1.2: Hypothetical emotional arcs corresponding to the amount of *joy* in three genres of literature.

in text in a structured way? The answers to these questions come in the next sections, which build upon research questions introduced in the next section.

1.2.2 Linguistic Structure of Emotion

Studies[5] following the research direction first articulated by Anderson and McMaster build upon certain simplifications regarding emotional arcs. Particularly, it is assumed that a mere counting of words associated with emotions using a dictionary has enough credibility to any analysis based on emotional arcs.

[5]Including my own work.

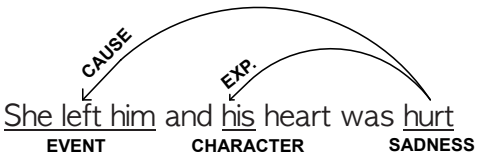


Figure 1.3: Example of structure of emotion. *Exp.* stands for experiencer or a person who is experiencing an emotion.

Although we shall see in Section 3 that these simplifications are not necessarily erroneous, they are, in fact, limiting. The biggest limitation imposed by emotion arcs is the crude granularity of aggregated emotion information, as this information spreads over all characters, narration, and descriptions. In other words, it is almost impossible to tell what the emotion arcs are really "made of".

One way to address this problem is by going from high level (emotion arc) to low level (linguistic structure). Consider Figure 1.3. One can see that the sentence depicts an instance of an *emotion* (sadness conveyed by "hurt") *experienced* by a *character* (expressed via "his") and this emotion has a *cause* corresponding to an *event* ("She left him").

The structure depicted on Figure 1.3 is a typical structure of an emotion (Russell and Barrett, 1999). As one may guess, representing every text with such a structure is not a trivial

task. One could imagine situations where the cause of the emotion is not articulate or where the emotion itself could not be reduced to a single word or phrase. However, the main question is why this type of analysis would be interesting? The answer is two-fold. On the one hand, if the analysis of emotional arcs gives a high-level impression about the emotion trend followed in the text, structural analysis aims at deeper understanding of depicted emotions. On the other hand, structural analysis alone does not lend itself to better interpretation of the role of emotions in literature. In other words, knowing the structure of emotions in text, we still need to come up with a way to use this knowledge for a high-level analysis.

These considerations bring forward the second research question I address in my thesis:

Research Question 2

Which parts of the emotion structure are the most difficult for the classification task and are the extraction tasks of different parts of emotion structure interrelated?

To that end, I will explore how difficult it is to 1) first annotate a literary text with emotions in a structured way, and

2) predict the relevant parts in the structure using computational modelling.

As we shall see in Section 4, the answer is not straightforward. I will discuss my main finding regarding these research questions later in the thesis. However, I would like to mention now that the search for answers to these questions will result in an interesting by-product that will turn into a research direction of its own. I discuss this matter in the next section.

1.2.3 Emotion-Informed Networks of Characters

One observation made by me while addressing the Research Question 2 is that informing the classifier about an emotion cue, simplifies the task of finding the experiencer of emotion. In other words, both are closely connected on a text level suggesting that they should be modeled jointly. Another observation is that literary characters show interesting behavior (or are depicted by authors in such a way), when it comes to emotion. Specifically, I observed that in many cases *characters are experiencing emotions because of other characters*. Consider the examples below:

Examples

1) "This man was a great bugbear to Fanny . . ." [FEAR]
2) "The mention of the widow singularly offended him . . ." [ANGER]
3) ". . . his mother never could be happy there without him." [JOY]

The textual spans corresponding to emotion (given in square brackets) cause are highlighted with orange, and spans corresponding to emotion experiencer with blue. It can be seen that in all of the situations one character is experiencing an emotion because of the other character. One may speculate from here that character relationships are an important tool the author uses to make a story progress. In fact, it has been argued that character interactions alone are able to move the plot forward (Booth, 2012, p. 321), and that "drama refers to interaction between characters, not conflict within a character. . ." (Bird, 2016, p. 64). This motivates an interesting research direction, as focusing on character relationships adds a potentially useful (for a literary scholar or NLP engineer) analysis angle to the study of emotions in text.

One way to approach the problem of character relationship analysis is to cast it as a joint *network* and *emotion* analysis

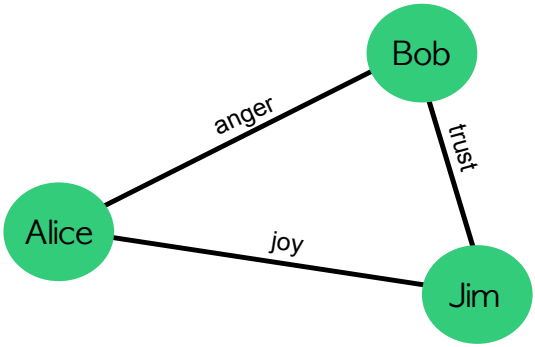


Figure 1.4: A hypothetical social network of characters in a book.

task. Figure 1.4 exemplifies such a task. We can see that there are three characters in a book, Alice, Bob, and Jim and their relationships could be summarized with emotions of *anger* (Alice–Bob), *trust* (Bob–Jim), and *joy* (Alice–Jim). In contrast to this toy example, literary texts usually depict dozens of interacting characters and, quite often, portray complicated relationship between them, which may or may not be singular (Alice and Bob's *anger* can alternate with *joy*, etc.). Being able to uncover such emotion-informed network structures could bring interesting insights into how emotion-based character interactions influence the plot. However, such an analysis is not possible without a robust approach to detecting character relationship. So, in my thesis, I will explore the problem of detecting non-propositional character relations and search for

the answer to the following research question:

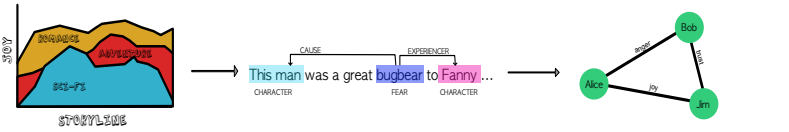
Research Question 3

Which model is suitable for capturing character relationships?

On the one hand, I will explore what classification algorithm suits the task the best. One the other hand, I will investigate what representation of characters is the most useful for the task. I will also try to understand the effect the number of modeled emotion classes has on the classification performance and explore a real-world implementation of the model void of some simplifications assumed in the exploration phase. Ultimately, with this questions I will try to understand *if the task is feasible*, and *what is needed* to make its results useful for a literary scholar or an NLP engineer.

1.3 Research Questions: Summary

To summarize, there are three major research questions I will address in this thesis. These questions are listed in a colored box below, as well as visualized in Figure 1.5.



High-level emotion analysis based on **emotion dictionary** and **emotion arcs**. Focus on **genre classification** using emotion arcs.

Low-level emotion analysis studying the **linguistic structure** of emotion. Focus on modeling emotion roles of experiencer, cause, target, and emotion.

Network-based analysis of the relationships between characters. Focus: characters and potential of using network features in **genre classification.**

Figure 1.5: Visual summary of research questions. The figure depicts the structure of my research regarding emotion arcs, genre classification, and network analysis.

Research Questions

1. Does emotion classification performed with dictionaries allow for the classification of genres?
2. Which parts of the emotion structure are the most difficult for the classification task and are the extraction tasks of different parts of emotion structure interrelated?
3. Which model is suitable for capturing character relationships?

I will start my thesis on a rather high-level investigating the extent to which emotion dictionaries are meaningful in the literary analysis. To that end, I set up a genre classification experiment with emotional arc features obtained with the help

of a dictionary.

Next, I will go down to the text (low) level and investigate the structure of emotion in text and inspect which parts of this structure are the most difficult for modelling.

Finally, I will shift the focus of analysis a bit up again and combine low-level character relationship analysis with a high-level network analysis, where characters are nodes in a graph, and edges between them are emotions that exist between them. I consider both binary (positive–negative) and non-binary (multiple emotions) character relationships.

I address these research questions in detail in Sections 3, 4, 5, and 6.

2 Background and Related Work

2.1 Foreword

This chapter deals with *emotion* and *sentiment* analysis in *computational literary studies*. Following the definition of Liu (2015), I define sentiment as a *positive* or *negative* feeling underlying the opinion. The term *opinion* in this sense is close to *attitude* in psychology and both sentiment analysis and opinion mining are often used interchangeably. Sentiment analysis is an area of computational linguistics that analyzes people's sentiments and opinions regarding different objects or topics. Sentiment analysis is primarily text-oriented, however, there are multimodal approaches as well (Soleymani et al., 2017).

Giving a definition to the concept of *emotion* is a challenging task. As Scherer (2005) puts it, "defining emotion is a notorious problem". Indeed, different methodological and conceptual approaches to dealing with emotions, lead to different definitions of thereof. However, the majority of emotion theorists agree

that emotions involve a set of expressive, behavioral, physiological, and phenomenological features (Scarantino, 2016). In this view, an emotion can be defined as "an integrated feeling state involving physiological changes, motor-preparedness, cognitions about action, and inner experiences that emerges from an appraisal of the self or situation" (Mayer et al., 2008).

Just as sentiment, emotions can be analyzed computationally. However, the goal of emotion analysis is to recognize the emotion, rather than sentiment, which makes it a more difficult task, as differences between emotions are subtler than between positive and negative.

Although sentiment and emotion analysis are different tasks, my review of the literature shows that the use of both terms is not always consistent. There are cases where researchers analyze only positive and negative aspects of a text but refer to their analysis as emotion analysis. Likewise, there are cases where researchers look into a set of subjective feelings including emotions but call it sentiment analysis. Hence, to avoid confusion, in this literature review, I use terms *emotion analysis* and *sentiment analysis* interchangeably. In most cases, I follow the terminology used by the authors of the papers I discuss (*i.e.*, if they call emotions sentiments, I do the same).

Finally, this chapter deals with sentiment and emotion analysis in the context of computational literary studies. Da (2019) defines computational literary studies as "the statistical repre-

sentation of patterns discovered in text mining fitted to currently existing knowledge about literature, literary history, and textual production." The most frequently applied approach within computational literary studies is *distant reading* (Moretti, 2005), which refers to the practice of running a textual analysis on a computer to yield quantitative results. In this literature review, I use these two terms interchangeably and when I refer to digital humanities as a field, I refer to those groups of researchers whose primary objects of study are texts.

2.2 Emotion in Psychology

The subject of emotion theories is vast and diverse. I refer the reader to Maria Gendron's paper (Gendron and Feldman Barrett, 2009) for a brief history of ideas about emotion in psychology. Here, I will focus on the three emotion views that are popular in computational analysis of emotions: Ekman's theory of basic emotions, Plutchik's wheel of emotion, and Russel's circumplex model.

2.2.1 Ekman's Theory of Basic Emotions

The basic emotion theory was first articulated by Silvan Tomkins (Tomkins, 1962) in the early 1960s. Tomkins postulated that each instance of a certain emotion is biologically similar to other instances of the same emotion or share a common trigger.

One of Tomkins' mentees, Paul Ekman, put in question the existing emotion theories that proclaimed that facial expressions of emotion are socially learned and therefore vary from culture to culture. Together with Sorenson and Friesen, Ekman endeavor on a field trip to other countries to challenge this view (Ekman et al., 1969). The outcome of their large-scale study led to the conclusion that facial displays of fundamental emotions are not learned but innate, and therefore are universal across the nationalities. However, there are culture-specific prescriptions about how and in which situations emotions are displayed.

Based on the observation of facial behavior in early development or social interaction, Ekman's theory also postulates that emotions should be considered discrete categories Ekman (1993), rather than continuous. Though this view allows for conceiving of emotions as having different intensities, it does not allow emotions to blend and leaves no room for more complex affective states in which individuals report the co-occurrence of like-valenced discrete emotions (Barrett, 1998). This and other theory postulates were widely criticized and disputed in literature (Russell, 1994; Russell et al., 2003; Gendron et al., 2014; Barrett, 2017).

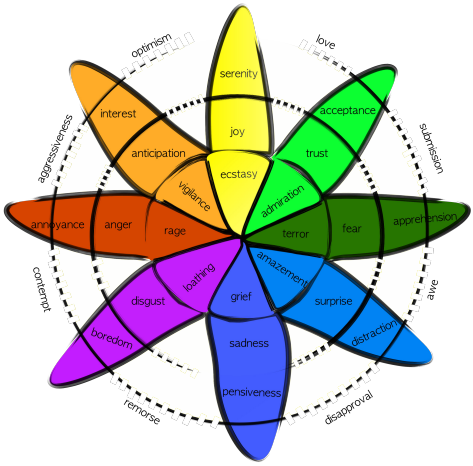


Figure 2.1: Plutchik's wheel of emotions. Depiction is my own work following the original Plutchik's depiction.

2.2.2 Plutchik' Wheel of Emotions

Another influential model of emotions was proposed by Robert Plutchik in the early 1980s (Plutchik, 1991). Important differences of Plutchik's theory from Ekman's theory is that apart from a small set of basic emotions, all other emotions are mixed and derived from the various combinations of basic ones. He further categorized these other emotions into the primary dyads (very likely to co-occur), secondary dyads (less likely to co-occur) and tertiary dyads (co-occur seldom).

In order to represent the organization and properties of the emotions as they were defined by his theory, Plutchik pro-

posed a structural model of emotions, which he called a *multidimensional model of emotions* that is more known today as *Plutchik's wheel of emotions*. The wheel (Figure 2.1) is constructed in the fashion of a color wheel, with similar emotions placed closer together and opposite emotions 180 degrees apart. The intensity of an emotion in the wheel depends on how far from the center a part of a petal is, *i.e.*, emotions become less intense the further they are from the center of the wheel. Essentially, the wheel is constructed from eight basic bipolar emotions: *joy* versus *sorrow*, *anger* versus *fear*, *trust* versus *disgust*, and *surprise* versus *anticipation*. The blank spaces between the leaves are so-called primary dyads – emotions that are mixtures of two of the primary emotions.

The wheel model of emotions proposed by Plutchik had a great impact on the field of affective computing being primarily used as a basis for emotion categorization in emotion recognition from text (Cambria et al., 2012; Kim et al., 2012; Suttles and Ide, 2013; Borth et al., 2013; Abdul-Mageed and Ungar, 2017). However, some postulates of the theory are criticized: for example, there is no empirical support for the wheel structure (Smith and Schneider, 2009); another criticism is that Plutchik's model of emotion does not explain the mechanisms by which love, hate, relief, pride, and other everyday emotions emerge from the "basic" emotions, nor does it provide reliable measurements of these emotions Richins (1997).

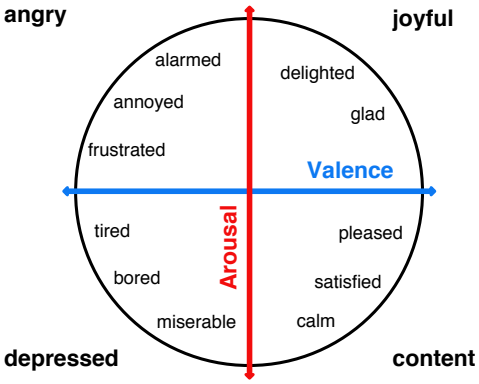


Figure 2.2: Russel's circumplex model.

2.2.3 Circumplex Model of Emotion by Russel

Attempts to overcome the shortcomings of basic emotions theory and its unfitness for clinical studies led researchers to suggesting various dimensional models, the most prominent of which is the circumplex model of affect proposed by Russell (1980). The word "circumplex" in the name of the model refers to the fact that emotion episodes do not cluster at the axes but at the periphery of a circle (Figure 2.2). At the core of the circumplex model is the notion of two dimensions plotted on a circle along horizontal and vertical axes. These dimensions are valence (how pleasant or unpleasant one feels) and arousal (the degree of calmness or excitement). The number of dimensions is not strictly fixed and there are adaptations of the model that incorporate more dimensions, as the Valence-

Arousal-Dominance model that adds an additional dimension of dominance, the degree of control one feels over the situation that causes an emotion (Bradley and Lang, 1994).

Essentially, by moving from discrete categories to a dimensional representation, the researchers are able to account for subjective experiences that do not fit nicely the isolated non-overlapping categories. Accordingly, each affective experience can be depicted as a point in a circumplex that is described by only two parameters – valence and arousal – without need for labeling or reference to folk emotion concepts (Russell, 2003). However, the strengths of the model turned out to be its weaknesses: For example, it is not clear if there are basic dimensions in the model (Larsen and Diener, 1992) and what to do with qualitatively different events of fear, anger, embarrassment and disgust that fall in identical places in the circumplex structure (Russell and Barrett, 1999). Despite these shortcomings, the circumplex model of affect is widely used in psychologic and psycholinguistic studies (*e.g.,* Sharar et al. (2016), Tseng et al. (2014), Baloglu and Love (2005)). In computational linguistics, the circumplex model is applied when the interest is in continuous measurements of valence and arousal rather than in the specific discrete emotion categories.

2.3 Emotion in Literary Studies

There were periods in literary and art theories when scholars disregarded the importance of the aesthetic and affective dimension of literature (Sætre et al., 2014). However, the affective turn taken by a wide range of disciplines in the past two decades – from political and sociological sciences to neurosciences to media studies – have refueled the interest of literary critics in human affects and sentiments.

As I said in Section 1.1, there is a strong movement in literary studies towards accepting emotion as an integral part of fictional texts. However, one might be challenged to define the specific way in which emotions come into play in the text. The exploration of this problem is presented by van Meel (1995). Underpinning the centrality of human destiny, hopes, and feelings in the themes of many artworks – from painting to literature – van Meel explores how emotions are involved in the production of arts. Pointing out to big differences between the two media in their possibilities to depict human emotions (paintings convey nonverbal behavior, such as facial expressions or gestures, directly, but lack temporal dimension that novels have and use to describe emotions), van Meel provides an analysis of the nonverbal descriptions used by the writers to convey emotion behavior of the characters. Description of visual characteristics, van Meel speculates, responds to a fundamental need of a reader to build an image of a person and

her behavior. Moreover, nonverbal descriptions add important information, which can in some cases play a crucial hermeneutical role, as in Kafka's *Der Prozess*, where the fatal decisions for K. are made clear by gestures rather than words. For example, his verdict is not announced, but is implied by the judge who refuses a handshake. The same applies to his death sentence that is conveyed to him by his executioners playing with a butcher's knife above his head.

A hermeneutic approach through the lenses of emotions is presented by Kuivalainen (2009), and provides a detailed analysis of linguistic features that contribute to characters' emotion involvement in Mansfield's prose. The study shows how, through the extensive use of adjectives, adverbs, deictic markers, and orthography, Mansfield steers the reader towards the protagonist's climax. Subtly shifting between psycho-narration[1] and free indirect discourse, Mansfield is making use of evaluative and emotive descriptors in psycho-narrative sections, often marking the internal discourse with dashes, exclamation marks, intensifiers, and repetition, which triggers an emotion climax. Various deictic features introduced in the text are used to pinpoint the source of emotions in the text, which helps in creating a picture of characters' emotion world. Verbs (especially, in present tense), adjectives, and adverbs serve the same goal in Mansfield's prose, to describe the emotion world

[1]Description of character consciousness.

of characters. Going back and forth from psycho-narration to free indirect discourse provides Mansfield with a tool to point out the significant moments in the protagonists' lives and draw a separation between characters and narration.

Both van Meel's and Kuivalainen works, separated from each other by more than a decade, underpin the importance of emotions in the interpretation of characters' traits, hopes, and tragedy.

Other authors find these connections as well. For example, Barton (1996) proposes instructional approaches to teach school-level readers to interpret emotions of the characters and use this information for story interpretation.

Van Horn (1997) shows that understanding characters emotionly or trying to help them with their problems made reading and writing more meaningful for middle school students.

Emotions in text are often conveyed with emotion-bearing words (Johnson-Laird and Oatley, 1989). At the same time their role in the creation and depiction of emotion should not be overestimated. That is, saying that someone looked angry or fearful or sad, as well as directly expressing characters emotions are not the only ways the authors resort to when building believable fictional space filled with characters, action, and emotions. In fact, many novelists strived to express emotions indirectly by way of figures of speech or catachresis (Miller, 2014), first of all, because emotion language can be ambiguous

and vague, and, second, to avoid any allusions to Victorian emotionism[2] and pathos.

2.4 Emotion in Natural Language Processing

2.4.1 Emotion Analysis in Computational Linguistics

In this section, I give an overview of the methods of emotion recognition from text. This section is limited to an overview of emotion analysis methods in natural language processing and is an important prerequisite for understanding the upcoming discussions in the main part of my thesis.

Although emotion recognition from text is a relatively new task, there are various approaches ranging from simple lexicon matching (Dodds et al., 2011), to deep learning methods (*e.g.*, Abdul-Mageed and Ungar, 2017), both supervised and unsupervised.

One of the earliest works on emotion detection from text is Alm et al. (2005) that uses supervised machine learning to classify sentences from children books into emotion categories using a large number of linguistic features, such as exclamation marks, ratio of adjectives, nouns, and verbs, direct speech. Aman and Szpakowicz (2007) is another early emotion anno-

[2]The thesis that emotions are in some manner essential to morality (Richard, 2009).

tation and classification study that uses lexical features in a bag-of-words fashion to classify sentences from a blog corpus into a set of Ekman's emotions.

The task of emotion classification of text is challenging. One possible reason for that is that the text should be classified in a larger number of classes. In most cases, it is conventional to use emotion classes from existing emotion theories as reference categories. That means that the task of emotion classification stretches beyond binary classification (e.g., positive vs. negative as in sentiment analysis). Ghazi et al. (2010) propose one way to tackle this problem with hierarchical classification, which first classifies the text as emotion or not emotion, and then performs a multiclass categorization on emotion text.

Emotions can have strong linguistic markers that define the tone of the text (Johnson-Laird and Oatley, 1989). This warrants different approaches to emotion detection from text that rely on linguistically rich features. For example, Neviarouskaya et al. (2009) propose rules to formulate how nouns, verbs or adjectives dominate the emotion in the corresponding sentence and combine those with modifiers. Gao et al. (2014) use dependency based features for classification, including negations. Agrawal and An (2012) propose an unsupervised framework for emotion extraction from sentences based on semantic relatedness between words and various emotion concepts.

As text classification in general, the array of methods seen

for emotion classification can be divided into rule-based methods and machine learning, which I discuss in the following.

Rule-based Algorithms

Rule-based text classification typically builds on top of lexical resources of emotionly charged words. These dictionaries can originate from crowdsourcing or expert curation. Examples include WordNetAffect (Strapparava and Valitutti, 2004) and SentiWordNet (Esuli and Sebastiani, 2006), both of which stem from expert annotation. Partly built on top of them is the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013), which uses the eight basic emotions from Plutchik classification. Warriner et al. (2013) use crowdsourcing to assign values of valence, arousal, and dominance (Russell, 1980). Another related category of lexical resources which has been used for emotion analysis is concreteness and abstractness (Köper et al., 2017). Brysbaert et al. (2014) publish a lexicon based on crowdsourcing, where the task was to assign a rating from 1 to 5 of the concreteness of 40,000 words. Similarly, Köper and im Walde (2016) automatically generate affective norms of abstractness, arousal, imageability, and valence for 350,000 lemmas in German. The Linguistic Inquiry and Word Count (LIWC) is a set of 73 lexicons (Pennebaker et al., 2007), built to gather aspects of lexical meaning regarding psychological tasks. Dictionary and rule-based approaches are particularly

common in the field of digital humanities due to their transparency and straightforward use.

Machine Learning

A performance improvement over dictionary lookup has been observed with supervised learning. Common features, depending on the text type, include word n-grams, character n-grams, word embeddings, affect lexicons, negation, punctuation, emoticons, or hashtags (Mohammad, 2012a). This feature representation is then usually used as input to feed classifiers such as naive Bayes, SVM, MaxEnt and others to predict the relevant emotion category (Alm et al., 2005; Aman and Szpakowicz, 2007). Similar to the paradigm shift in sentiment analysis, from feature-based modelling to deep learning, state-of-the-art models for emotion classification are often based on neural networks. Schuff et al. (2017) applied models from the classes of CNN, biLSTM, and LSTM and compare them to linear classifiers (SVM and MaxEnt), where the biLSTM show best results with the most balanced precision and recall. Abdul-Mageed and Ungar (2017) claim the highest F_1 following Plutchik's emotion model with gated recurrent unit networks (Chung et al., 2015). One approach to tackle sparsity of datasets is transfer learning; to make use of similar resources and then transfer the model to the actual task. A recent successful example for this procedure is Felbo et al. (2017) who present a

neural network model trained on emoticons which is then transferred to different downstream tasks, namely the prediction of sentiment, sarcasm, and emotions.

A recent shared task on implicit emotion detection (IEST, Klinger et al., 2018) showed that the majority of participants built systems on top of deep learning architectures, similarly to participants of the emotion intensity shared tasks in previous years Mohammad and Bravo-Marquez (2017); Mohammad et al. (2018). Therefore, we can conclude that the paradigm shift to deep learning has reached the field of emotion analysis.

2.4.2 Emotion Analysis in Digital Humanities

With this section, I proceed to an overview of the existing body of research on computational analysis of emotion and sentiment in computational literary studies.

Emotion Classification

A straightforward approach to sentiment and emotion analysis is phrasing them as a text classification (Liu, 2015). A fundamental question of such a classification is how to find the best features and algorithms to classify the data (sentences, paragraphs, entire documents) into predefined classes. When applied to literature, such a classification may be of use for grouping different literary texts in digital collections based on the emotion properties of the stories. For example, books or

poems can be grouped based on the emotions they convey or based on whether they have happy endings or not.

Classification based on emotions Barros et al. (2013) aim at answering two research questions: 1) Is the classification of Quevedo's works proposed by the literary scholars consistent with the sentiment reflected by the corresponding poems? and 2) Which learning algorithms are the best for the classification? To that end, they perform a set of experiments on the classification of 185 Francisco de Quevedo's poems that are divided by literary scholars into four categories and that authors map to emotions of joy, anger, fear, and sadness. Using terms "joy", "anger", "fear", and "sadness" as a point of reference, Barros et al. construct the list of emotion words by looking up the synonyms of English emotion words and adjectives associated with these four emotions and translating them to Spanish. Each poem is converted into a vector where each item is a normalized count of words relating to a certain emotion. The experiments with different algorithms show the superiority of decision trees achieving accuracy of almost 60%. However, this result is biased by unbalanced distribution of classes. To avoid the bias, Barros et al. apply a resampling strategy that leads to a more balanced distribution and repeat the classification experiments. After resampling, the accuracy of decision trees in a 10-fold cross validation achieves 75,13%,

thus improving over the previous classification performance. Based on these results the authors conclude that a meaningful classification of the literary pieces based only on the emotion information is possible.

Reed (2018) offers a proof-of-concept for performing sentiment analysis on a corpus of the twentieth-century American poetry. Specifically, Reed analyzes the expression of emotions in the poetry of the Black Arts Movement of the 1960s and 1970s. The paper describes the project "Measured Unrest in the Poetry of the Black Arts Movement" whose goal is to understand 1) how the feelings associated with injustice are coded in terms of race and gender, and 2) what can sentiment analysis show us about the relations between affect and gender in poetry. Reed notes that surface affective value of the words does not always align with their more nuanced affective meaning shaped by poetic, social, and political contexts.

Yu (2008) explores what linguistic patterns characterize the genre of sentimentalism in early American novels. To that end, they construct a collection of five novels from the mid-nineteenth century and annotate the emotionity of each of the chapters as "high" or "low". The respective chapters are then classified using support-vector machines and naïve Bayes classifiers as highly emotion or the opposite. The results of the evaluation suggest that arbitrary feature reduction steps such as stemming and stopword removal should be taken very care-

fully, as they may affect the prediction. For example, Yu shows that no stemming leads to better classification results. A possible explanation is that stemming conflates and neutralizes a large number of discriminative features. The author provides an example of such a conflation with words "wilderness" and "wild." – while the latter can appear anywhere in the text, the former one is primarily encountered in the chapters filled with emotions.

Classification of happy ending vs. non-happy endings Zehe et al. (2016) argue that automatically recognizing a happy ending as a major plot element could help to better understand a plot structure as a whole. To show that this is possible, they classify 212 German novels written between 1750 and 1920 as having happy or non-happy ending. A novel is considered to have a happy ending if the situation of the main characters in the novel improves towards the end or is constantly favorable. The novels were manually annotated with this information by domain experts. For feature extraction, the authors first split each novel into n segments of the same length. Then they calculate sentiment values for each of the segments by counting the occurrences of words that appear in the respective segment and are found in the German version of NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013) and dividing this number by the number of dictionary words in the

segment. Finally, they calculate the sentiment score for the sections, by taking the average of all sentiment scores in the segments that are part of the section. These steps are then followed by classification with a support-vector machines algorithm and the F_1 score of 0.73, which the authors consider a good starting point for future work.

Genre and Story-type Classification

The papers I discussed so far focus on understanding the emotion associated to units of texts. This extracted information can further be used for downstream tasks and also for downstream evaluations. I discuss the following downstream classification cases here. The papers in this category use sentiment and emotion features for a higher-level classification, namely story-type clustering and literary genre classification. The assumption behind these works is that different types of literary text may show different composition and distribution of emotion vocabulary, and thus, can be classified based on this information. The hypothesis that different literary genres convey different emotions stems from the common knowledge: we know that horror stories instill fear, mysteries evoke anticipation and anger, while romances are filled with joy and love. However, the task of automatic classification of these genres is not always that straight-forward and reliable, as we will see in this section.

Story-type clustering Similarly to Zehe et al., Reagan et al. (2016) are interested in automatically understanding a plot structure as a whole, not limited to a book ending. The inspiration for their work comes from Kurt Vonnegut's lecture on emotion arcs of stories[3]. Reagan et al. test the idea that the plot of each story can be plotted as an *emotion arc*, i.e. a time series graph, where the x-axis represents a time point in a story, and the y-axis represents the events happening to the main characters, which can be favorable (peaks on a graph) or unfavorable (troughs on a graph). As Vonnegut puts it, the stories can be grouped by these arcs and the number of such grouping is limited. To test this idea, Reagan et al. collect 1,327 most popular books from the Project Gutenberg[4]. Each book is then split into segments for which sentiment scores (happy vs. sad) are calculated and compared. The results of the analysis show support for six emotion patterns that are shared between subgrouping of the corpus:

- Rise ("Rags to riches"): the arc starts at low point and steadily increases towards the end;
- Fall ("Tragedy"): the arc starts at high point and steadily decreases towards the end;

[3] As available on Web Archive as of March 2018 `https://web.archive.org/web/20100326094804/http:/www.laphamsquarterly.org:80/voices-in-time/kurt-vonnegut-at-the-blackboard.php?page=all`

[4] `https://www.gutenberg.org/`

- Fall-rise ("Man in a hole"): the arc drops in the middle of the story but increases towards the end;
- Rise-fall ("Icarus"): the arc hits the high point in the middle of the story decreases towards the end;
- Rise-fall-rise ("Cinderella"): the arc fluctuates between high and low points but ends with an increase;
- Fall-rise-fall ("Oedipus"): the arc fluctuates between high and low points but ends with a decrease.

Additionally, Reagan et al. examine the downloads for all the books that are most similar and find that "Icarus", "Oedipus", and "Man in a hole" arcs are the three most popular emotion arcs among readers.

Genre classification A study by Samothrakis and Fasli (2015) is similar in spirit to the work by Reagan. Samothrakis and Fasli examine the hypothesis that different genres have clearly different emotion patterns to reliably classify them with machine learning. To that end, they collect works of genres *mystery, humor, fantasy, horror, science fiction* and *western* from the Project Gutenberg. Using WordNet-Affect (Strapparava and Valitutti, 2004) to detect emotion words as categorized by Ekman's fundamental emotion classes, they calculate an emotion score for each sentence in the text. Each work is then

transformed into six vectors, one for each basic emotion. A random forest classifier achieves classification accuracy of 0.52. This is significantly higher than a random baseline, which allows the authors to conclude that such a classification is feasible.

The study by Henny-Krahmer (2018) pursues two goals: first, to test whether different subgenres of Spanish American literature differ in the degree and kind of emotionity, and second, whether emotions in the novels are expressed in the direct speech of characters or in narrated text. To that end, they conduct a subgenre classification experiment on a corpus of Spanish American novels using sentiment values as features. To answer the first question, each novel is split into five segments and for each sentence in the segment the emotion score (polarity values + Plutchik's basic emotions) is calculated using SentiWordNet (Baccianella et al., 2010) and NRC (Mohammad and Turney, 2013) dictionaries. The classifier achieves an average F_1 of 0.52, which is higher than the most-frequent class baseline and, hence, provides a support for emotion-based features in subgenre classification. The analysis of feature importance shows that the most salient features come from the sentiment scores calculated from the direct speech of the characters, and that novels with higher values of positive speech are more likely to be sentimental novels.

There are some limitations to the studies presented in this

section. On the one hand, it is questionable how reliable is coarse emotion scoring that takes into account only presence or absence of words found in specialized dictionaries and overlook negations and modifiers that can either negate an emotion word or increase/decrease its intensity. On the other hand, a limited view of the emotion content as a sum of emotion bearing words reserves no room for qualitative interpretation of the texts – it is not clear how can one distinguish between emotion words used by the author to express his/her sentiment, between words used to describe character's feelings, and emotion words that characters use to address or describe other characters in a story.

Temporal Change of Sentiment

The papers that I have reviewed so far approach the problem of sentiment and emotion analysis as a "static" classification task. However, applications of sentiment analysis are not only limited to classification. In other fields, for example computational social sciences, sentiment analysis can be used for analyzing political preferences of the electorate or for mining opinions about different products or topics. Similarly, several digital humanities studies incorporate sentiment analysis methods in a task of mining sentiments and emotions of people who lived in the past. The goal of these studies is not only to recognize sentiments, but also to understand how they were formed.

Topography of emotions Heuser et al. (2016) start with a premise that emotions occur at a specific moment in time and space, thus making it possible to link emotions to specific geographical locations. Consequently, having such information at hand, one can understand which emotions are hidden behind certain landmarks. As a proof-of-concept, Heuser et al. build an interactive map of emotions in Victorian London[5] where each location is tagged with emotion labels. To construct a corpus for their analysis, Heuser et al. collect a large corpus of English books from eighteenth and nineteenth centuries and extract 383 geographical locations of London that have at least ten mentions each. The resulting corpus includes 15,000 passages, each of which has a toponym in the middle and 100 words directly preceding and following the location mention. The data is then given to annotators who are asked to define whether each of the passages expressed happiness or fear, or neutral. The same data is also analyzed by a custom sentiment analysis program that would assign each passage one of these emotion categories.

Some striking observations are made with regard to the data analysis. First, there is a clear discrepancy between fiction and reality – while toponyms from the West End with its Westminster and the City are over-represented in the books, the same does not hold true for the East End with its Tower Ham-

[5] `https://www.historypin.org/en/victorian-london/`

lets, Southwark, and Hackney. Hence, there is less information about emotions of these particular London locations. Another striking detail is that the resulting map is dominated by the neutral emotion. Heuser et al. argue that this has nothing to do with the absence of emotions but rather stems from the fact that emotions tend to be silenced in public domain, which influenced the annotators decision.

The space and time context are also used by Bruggmann and Fabrikant (2014) who model sentiments of the Swiss historians towards places in Switzerland in different historical periods. As the authors note, it is unlikely that a historian will directly express attitudes towards certain toponyms, but it is very likely that words they use to describe those can bear some negative connotation (e.g., Cholera, death). Correspondingly, such places should be identified as bearing negative sentiment by sentiment analysis tool. Additionally, they study the changes of sentiment towards a particular place over time. Using the *General Inquirer* (GI) lexicon (Stone et al., 1968) to identify positive and negative terms in the document, they assign each document a sentiment score by summing up the weights of negative and positive words and normalizing them by the document length. The authors conclude that the results of their analysis look promising, especially regarding negatively scored articles. However, the authors find difficulties interpreting positively ranked documents, which may be due to the fact that

negative information is more salient.

Tracking sentiment Other papers in this category link sentiment and emotion to certain groups, rather than geographical locations. The goal of these studies is to understand how sentiment within and towards these groups was formed.

Taboada et al. (2006, 2008) aim at tracking the literary reputation of six authors writing in the first half of the twentieth century. The research questions raised in the project are how the reputation is made or lost, and how to find correlation between what is written about the author and his/her work to the author's reputation and subsequent canonicity. To that end, the project's goal is to examine critical reviews of six authors writing and map information contained in the critical texts to the author's reputation. The material they work with include not only reviews, but also press notes, press articles, and letters to editors (including from the authors themselves). For the pilot project with Galsworthy and Lawrence they have collected and scanned 330 documents (480,000 words). The documents are tagged for the parts-of-speech and relevant words (positive and negative) are extracted using custom-made sentiment dictionaries. The sentiment orientation of rhetorically important parts of the texts is then measured.

Chen et al. (2012) aims to understand personal narratives of Korean "comfort women" who had been forced into sexual

slavery by Japanese military during World War II. Adapting the WordNet-Affect lexicon (Strapparava and Valitutti, 2004), Chen et al. build their own emotion dictionary to spot emotion keywords in women's stories and map the sentences to emotion categories. By adding variables of time and space, Chen et al. provide a unified framework of collective remembering of this historical event as witnessed by the victims.

Sentiment recognition in historical texts Other papers put emphasis not so much on the sentiments expressed by the writers of the past, but rather on methodology of sentiment detection from old texts.

Marchetti et al. (2014) and Sprugnoli et al. (2016) present the integration of sentiment analysis in ALCIDE (Analysis of Language and Content In a Digital Environment) project[6]. The sentiment analysis module is based on WordNet-Affect, SentiWordNet (Baccianella et al., 2010) and MultiWordNet (Pianta et al., 2002). Each document is assigned with a polarity score by summing up the words with prior polarity and dividing by the number of words in the document. A positive global score leads to a positive document polarity and a negative global score leads to negative document polarity. The overall conclusion of their wokr is that the assignment of a polarity in the historical domain is a challenging task largely due

[6] `http://celct.fbk.eu:8080/Alcide_Demo/`

to lack of agreement on polarity of historical sources between human annotators.

Challenged by the problem of applicability of existing emotion lexicons to historical texts, Buechel et al. (2017) propose a new method of constructing affective lexicons that would adapt well to the German texts written up to three centuries ago. In their study, Buechel et al. use the representation of affect based on the Valence-Arousal-Dominance model (an adaptation of Russel's circumplex model, see Section 2.2.3). Presumably, such a representation provides a finer-grained insight into the literary text (Buechel et al., 2016), which is more expressive than discrete categories, as it quantifies the emotion along three different dimensions. As a basis for the analysis, they collect German texts from the *Deutsches Textarchiv*[7] written between 1690 and 1899. The corpus is split into seven slices, each spanning 30 years. For each slice they compute word similarities and obtain seven distinct emotion lexicons, each corresponding to specific time period. This allows, the authors argue, tracing the shift in emotion association of words over time.

Finally, Leemans et al. (2017) aim to trace historical changes in emotion expressions and to develop methods to trace these changes in a corpus of 29 Dutch language theatre plays written between 1600 and 1800. Extending the Dutch version of

[7] `http://www.deutschestextarchiv.de/`

Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007) with historical terms, the authors are able to increase the recall of emotion recognition with a dictionary. In addition, they develop a fine-grained vocabulary mapping body terms to emotions and show that a combination of LIWC and their lexicon lead to the improvement in the emotion recognition.

Character Network Analysis and Relationship Extraction

The papers reviewed above address sentiment analysis of literary texts mainly on a document level. This abstraction is warranted if the goal is to get an insight into the distribution of emotions in a corpus of books. However, emotions depicted in books do not exist in isolation, but are associated with characters, who are at the core of any literary narrative (Ingermanson and Economy, 2009). Characters are a central feature of most kinds of fiction (Eder et al., 2010, p. 3) and they have been extensively studied in the academic literature (Jannidis, 2008; Phelan, 1989; Palmer, 2004).

This leads us to a question of what sentiment and emotion analysis can tell us about the characters. How emotional are they? And what role do emotions play in their interaction?

Character relationships have been analyzed in computational linguistics from a graph theoretic perspective, particularly using social network analysis (Agarwal et al., 2013; Elson et al.,

2010). Fewer works, however, address the problem of modelling character relationships in terms of sentiment. Below I provide an overview of several papers that propose the methodology for extracting this information.

Sentiment dynamics between characters Some studies are concerned with automatic methods for analyzing sentiment dynamics between plays' characters. The goal of the study by Nalisnick and Baird (2013a) is to track the emotion trajectories of interpersonal relationship. The structured format of a dialog allows them to identify who is speaking to whom, which makes it possible to mine character-to-character sentiment by summing the valence values of words that appear in the continuous direct speech and are found in the lexicon (Nielsen, 2011) of affective norms. The extension (Nalisnick and Baird, 2013b) of the previous research from the same authors introduces the concept of a "sentiment network", a dynamic social network of characters. Changing polarities between characters are modeled as edge weights in the network. Motivated by the desire to explain such networks in terms of general sociological model, Nalisnick and Baird test if Shakespeare's plays obey the Structural Balance Theory by Marvel et al. (2011) that postulates that a friend of a friend is also your friend. Using the procedure proposed by Marvel et al. on their Shakespearean sentiment networks, Nalisnick and Baird test if they

can predict how a play's characters will split into factions using only information about the state of the sentiment network after Act II. The results of their analysis are varied and do not provide adequate support for the Structural Balance Theory as a benchmark for network analysis in Shakespeare's plays. One reason for that, as the authors state, is inadequacy of their shallow sentiment analysis methods that cannot detect such elements of speech as irony and deceit that play pivotal role in many literary works.

Character analysis and character relationships Elsner (2012, 2015) aims at answering the question of how to represent a plot structure for summarization and generation tools. To that end, Elsner presents a "kernel" for comparing novelistic plots at the level of character interactions and their relationships. Using sentiment as one of the characteristics of a character, Elsner demonstrates that the kernel approach leads to meaningful plot representation that can be used for a higher-level processing.

Barth et al. (2018) develop a character relation analysis tool *rCAT* with the goal of visualization and analysis of character networks in literary text. The tool implements a distance parameter (based on token space) for finding pairs of interacting characters. In addition to the general context words that characterize each pair of characters, the tool provides an emotion filter to restrict character relationship analysis to emotions

only.

A tool presented by Jhavar and Mirza (2018) provides a similar functionality: given an input of two character names from *Harry Potter* series, the EMOFIEL[8] tool identifies the emotion flow between a given directed pair of story characters. These emotions are identified using categorical (Plutchik, 1991) and continuous (Russell, 1980) emotion models.

Egloff et al. (2018) present an ongoing work on the Ontology of Literary Characters (OLC) that allows to capture and infer psychological characters' traits from their linguistic descriptions. The OLC incorporates the Ontology of Emotion (Patti et al., 2015) that is based on both Plutchik's and Hourglass's (Cambria et al., 2012) models of emotions. The ontology encodes thirty-two emotion concepts. Based on their natural language description, characters are attributed a psychological profile along the classes of Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The ontology links each of these profiles to one or more archetypal categories of *hero*, *anti-hero*, and *villain*. Egloff et al. argue that using the semantic connections of the OLC, it is possible to infer characters' psychological profiles and their role played in the plot.

Finally, a small body of work exists that focuses on mathematical modelling of character relationships. Rinaldi et al.

[8] `https://gate.d5.mpi-inf.mpg.de/emofiel/`

(2013) contribute a model that describes the love story between the Beauty and the Beast through ordinary differential equations. Zhuravlev et al. (2014) introduce a distance function to model the relationship between protagonist and other characters in two masochistic short novels by Ivan Turgenev and Sacher-Masoch. Borrowing some instruments from the literary criticism and using ordinary differential equations, Zhuravlev et al. are able to reproduce the temporal and spatial dynamics of the love plot in the two novellas more precisely than in had been done in previous research. Jafari et al. (2016) present a dynamic model describing the development of character relationships based on differential equations. The proposed model is enriched with complex variables that can represent complex emotions such as coexisting love and hate.

Other Types of Emotion Analysis

We have seen that sentiment analysis as applied to literature can be used for a number of downstream tasks, such as classification of texts based on the emotions they convey, genre classification based on emotions, and sentiment analysis on the historical domain. However, the application of sentiment analysis is not limited to these tasks. In this concluding part of the survey, I review some papers that do not formulate their approach to sentiment analysis as a downstream task. Often, the goal of these works is to understand how sentiments and

emotions are represented in literary text in general, and how sentiment or emotion content varies across specific documents or a collection of thereof with time, where time can be either relative to the text in question (from beginning to end) or to the historical changes in language (from past to present). Such information is valuable for gaining a deeper insight into how sentiments and emotions change over time allowing to bring forward new theories or shed more light onto existing literary or sociological theories.

Emotion flow analysis and visualization A set of authors aimed at visualizing the change of emotion content through texts or across time. One of the earliest works in this direction is a paper by Anderson and McMaster (1986) that starts from the premise that reading enjoyment stems in the affective tones of a text. These affective tones create a conflict, that can rise to climax through a series of crises, which is necessary for a work of fiction to be attractive to the reader. Using a list of 1,000 most common English words annotated with valence, arousal, and dominance ratings (Heise, 1965), they calculate the conflict score by taking the mean of the ratings for each word in a text passage. The more negative the score is, the higher the conflict is, and vice versa. Additionally, they plot conflict scores for each consecutive 100 words of a test story and provide qualitative analysis of the peaks. They argue that

a reader who has access to text would be able to find correlation between events in the story and peaks on the graph. However, the authors still stress that such interpretation remains dependent upon the judgement of the reader. Further, more practical, contributions by the authors are based on the same premises (Anderson and McMaster, 1982, 1993).

Alm and Sproat (2005) present the results of the emotion annotation task of 22 brothers Grimm's tales and evaluates patterns of emotion story development. They split emotions into *positive* and *negative* categories and divide each story into five parts for which aggregate frequency counts of combined emotion categories are computed. The resulting numbers are plotted on a graph that shows a wave-shaped pattern. From this graph, Alm and Sproat argue, one can see that the first part of the fairy tales is the least emotion, which is probably due to scene setting, while the last part shows an increase in positive emotions, which may signify the happy ending.

Two other studies by Mohammad (Mohammad, 2011, 2012b) focus on differences in emotion word density, as well as emotion trajectories, between books of different genres. Emotion word density is defined as a number of times a reader will encounter an emotion word on reading every X words. In addition, each text is assigned several emotion scores for each emotion that are calculated as a ratio of words associated with one emotion to the total number of emotion words occurring in a text. Both

metrics use the NRC Affective Lexicon to find occurrences of emotion words. They find that fairy tales have significantly higher *anticipation, disgust, joy* and *surprise* word densities, but lower *trust* word density when compared to novels.

A work by Klinger et al. (2016) is a case study in an automatic emotion analysis of Kafka's "Amerika" and "Das Schloss". The goal of the work is to analyze the development of emotions in both texts, as well as to provide a character-oriented emotion analysis that would reveal specific character traits in both texts. To that end, Klinger et al. develop German dictionaries of words associated with Ekman's fundamental emotions plus *contempt* and apply them to both texts in question to automatically detect emotion words. The results of their analysis for "Das Schloss" show a striking increase of *surprise* towards the end and a peak of *fear* shortly after start of chapter 3. In the case of "Amerika", the analysis shows that there is a decrease in *enjoyment* after a peak in chapter 4.

The work by Kakkonen and Galić Kakkonen (2011) aims at supporting the literary analysis of Gothic texts at the sentiment level. The authors introduce a system called SentiProfiler that generates visual representations of affective content in such texts and outlines similarities and differences between them, however, without considering the temporal dimension. The SentiProfiler uses WordNet-Affect to derive a list of emotion-bearing words that will be used for analysis. The resulting

sentiment profiles for the books are used to visualize the presence of sentiment in a particular document and to compare two different texts.

Miscellaneous In this section, I review studies that are different in goals and research questions from the papers presented in previous sections and do not constitute a category on their own.

Koolen (2018) claims that there is a bias among readers that put works by female authors on par with "women's books", which are, as stated by the author, tend to be perceived as of lower literary quality. She investigates how much "women's books" (here, romantic novels written by women) differ from novels that are perceived as literary (female and male-authored literary fiction). The corpus used in the study is the collection of European and North-American novels translated into Dutch. Koolen uses a Dutch version of the Linguistic Inquiry and Word Count (Boot et al., 2017), a dictionary that contains content and sentiment-related categories of words, to count the number of words from different categories in each type of fiction. Her analysis shows that romantic novels contain more positive emotions and words pertaining to friendship than in literary fiction. However, female-authored literary novels and male-authored ones do not significantly differ on any category.

Kraicer and Piper (2019) explore the women's place within contemporary fiction starting from the premise that there is a near ubiquitous underrepresentation and decentralization of women. As a part of their analysis, Kraicer and Piper use sentiment scores to look at social balance and "antagonism", *i.e.*, how different gender pairings influence positive and negative language surrounding the co-occurrence of characters (using the sentiment dictionary presented by Liu (2010) to calculate a sentiment score for a character pair). Having analyzed a set of 26,450 characters from 1,333 novels published between 2001 and 2015, the authors found that sentiment scores give little indication that the character's gender has an effect on the state of social balance.

Morin and Acerbi (2017) focus on larger-scale data spanning hundred thousand of books. The goal of their study is to understand how emotionity of written texts changed throughout the centuries. Having collected 307,527 books written between 1900 and 2000 from the Google Books corpus[9] they collect, for each year, the total number of case-insensitive occurrences of emotion terms that are found under positive and negative taxonomies of LIWC dictionary (Pennebaker et al., 2007). The main findings of their research show that emotionity (both positive and negative emotions) declines with time, and this

[9] `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`

decline is driven by the decrease in usage of positive vocabulary. Morin and Acerbi remind that the Romantic period was dominated by emotionity in writing, which could be the effect of a group of writers who wrote above the mean. If one assumes that each new writer tends to copy the emotion style of their predecessors, then writers at one point of time are disproportionally influenced by this group of above-the-mean writers. However, this trend does not last forever and, sooner or later, the trend reverts to the mean, as each writer reverts to a normal level of emotionity.

An earlier work (Bentley et al., 2014) written in collaboration with Acerbi provides a somewhat different approach and interpretation of the problem of the decline in positive vocabulary in English books of twentieth century. Using the same dataset and lexical resources (plus WordNet-Affect) Bentley et al. find a strong correlation between expressed negative emotions and the U.S. economic misery index, which is especially strong for the books written during and after the World War I (1918), the Great Depression (1935), and the energy crisis (1975). However, in the present study (Morin and Acerbi, 2017), the authors argue that the extent to which positive emotionity correlates with subjective well-being is a debatable issue. Morin and Acerbi provide more possible reasons for this effect, as well as detailed statistical analysis of the data, so I refer the reader to the original paper for more information.

3 Emotion and Literary Genres

3.1 Introduction[1]

In Section 1, I outlined several research directions pursued in this thesis. In particular, I speculated that emotions are important for literature and their study can benefit both literary scholars and NLP engineers. So, if emotion is important for literature, does it mean that it is a part of a text itself and can we detect these emotions using an emotion dictionary? If so, can we use the detected emotion for a genre classification? These considerations give rise to the following research question introduced in the beginning of this thesis:

[1]This chapter is an extension of Kim et al. (2017a).

Research Questions 1

Does emotion classification performed with dictionaries allow for the classification of genres?

To answer this question, I will use a dictionary to represent books via their emotional arcs. I will then set up two text classification experiments, one based on standard lexical features, another solely on *EmoArc* features. This way, it will become possible to measure the effect of emotional arcs features (and, consequently, dictionaries) on a downstream task, that is genre classification. In this section, I will explain in detail how I validate Research Question 1 on a sample of 2,000 fictional books from different literary genres.

There is little agreement between literary theorists about genre (Underwood, 2016). Generally genre is understood as a key means by which the many forms of literature and culture are categorized (Frow, 2015). Given the fact that there is no consensus on genre, these categorizations cannot be an absolute truth. However, in my work I follow the categorization of literary genres in the Brown corpus (Francis and Kucera, 1979) that divides fiction into *general fiction*, *adventures*, *mystery and detective fiction*, *romances*, *science fiction*, and *humor*.

3.2 Contributions

My main contributions related to this topic are as follows:

1. I carry out a large-scale analysis of emotion arcs on a sample of more than 2,000 books from Project Gutenberg[2]. My analysis covers the genres of *science fiction*, *adventure*, *humor*, *romantic fiction*, and *detective and mystery stories* that are defined in the Brown corpus.

2. I define two emotion-based models for genre classification based on the eight fundamental emotions defined by Plutchik (1991) – *fear*, *anger*, *joy*, *trust*, *surprise*, *sadness*, *disgust*, and *anticipation* (see Section 2.2.2). The first one is an emotion lexicon model based on the NRC dictionary (Mohammad and Turney, 2013). The second one is an emotion arc model that models the emotion development over the course of a story. I avoid the assumption of Reagan et al. (2016) that absence of happiness indicates fear or sadness.

3. I analyze the performance of the various models quantitatively and qualitatively. Specifically, I investigate how uniform genres are with respect to emotion developments and discuss differences in the importance of lexical units.

[2] `www.gutenberg.org`

Genre	Count
adventure	569
humor	202
mystery	379
romance	327
science fiction	542
$\sum$	2019

Table 3.1: Statistics for my Gutenberg genre corpus.

3.3 Data collection

I collect books from Project Gutenberg that match certain tags, namely those which correspond to the five literary genres found in the Brown corpus (Francis and Kucera, 1979): adventure (Gutenberg tag: "Adventure stories"), romance ("Love stories" and "Romantic fiction"), mystery ("Detective and mystery stories"), science fiction ("Science fiction"), and humor ("Humor"). All books must additionally have the tag "Fiction". I exclude books which contain one of the following tags: "Short stories", "Complete works", "Volume", "Chapter", "Part", "Collection". This leads to a corpus of 2113 stories. Out of these, 94 books (4.4 %) have more than one genre label. For simplicity, I discard these texts, which leads to the corpus of 2019 stories with the relatively balanced genre distribution as shown in Table 3.1.

3.4 Methods

3.4.1 Feature Sets

In the following, I describe the methods I use that aim to answer whether emotion classification with dictionaries allows for the genre classification. I will carry out classification with different feature sets. This will make it possible to see the difference between the features based on dictionaries and other features.

I consider three different feature sets: bag-of-words features (as a strong baseline), lexical emotion features, and emotion arc features.

Bag-of-words

An established strong feature set for genre classification, and text classification generally, consists of bag-of-words features. For genre classification, the generally adopted strategy is to use the n most frequent words in the corpus, whose distribution is supposed to carry more genre-specific rather than content- or domain-specific information. The choice of n varies across stylometric studies, from, *e.g.*, 1,000 (Sharoff et al., 2010) to 10,000 (Underwood, 2016). Here, I use $n = 5,000$ and refer to this feature set as BoW.

Emotion bag-of-words

My second feature set, EmoLex, is a filtered version of BoW, capturing lexically expressed emotion information. This feature aims to shed light on the question if emotion is an important part of a literary text. It consists of all words in the intersection between the corpus vocabulary and the NRC dictionary (Mohammad and Turney, 2013) which contains 4,463 words associated with 8 emotions. Thus, it incorporates the assumption that words associated with emotions reflect the actual emotion content (Bestgen, 1994). I do not take into account words from "positive"/"negative" categories or those that are not associated with any emotions. This model takes into account neither emotion labels nor position of an emotion expression in the text.

Emotion arc

The final feature set, EmoArc, in contrast to the lexical emotion features, takes into account both emotion labels and position of an emotion expression. It represents an emotion arc in the spirit of Reagan et al. (2016), but considers all of Plutchik's eight fundamental emotion classes. I split each input text into k equal-sized, contiguous segments S corresponding to spans of tokens $S = \langle t_n, \ldots, t_m \rangle$. I treat k as a hyper-parameter to be optimized.

I define a score $\text{es}(e, S)$ for the pairs of all segments S and each emotion e as

$$\text{es}(e, S) = \frac{c}{|D_e| \cdot |S|} \sum_{t_i \in S} \mathbf{1}_{t_i \in D_e} ,$$

where D_e is the NRC dictionary associating words with emotions, c is a constant set for convenience to the maximum token length of all texts in the corpus C ($c = \max_{S \in C} |S|$), and $\mathbf{1}_{t_i \in D_e}$ is 1 if $t_i \in D_e$ and 0 otherwise. This score, which makes the same assumption as the lexical emotion features, represents the number of words associated with emotion e per segment, normalized in order to account for differences in vocabulary size and book length.

The resulting features form an $8 \times k$ "emotion-segment" matrix (Figure 3.1) for each document that reflects the development of each of the eight emotions throughout the time-course of the narrative.

3.4.2 Models for Genre Classification

In the following, I discuss the use of the feature sets defined in Section 3.4.1 with classification methods to yield concrete models.

I use the two lexical feature sets, BoW and EmoLex, with a random forest classifier (RF, Breiman (2001)) and multilayer perceptron (MLP, Hinton (1989)). RF often performs

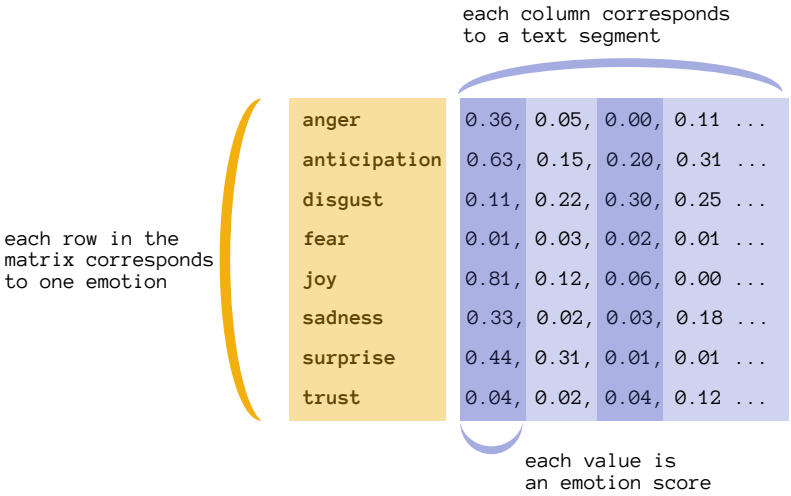


Figure 3.1: Example of emotion arc feature matrix.

well independent of chosen meta parameters (Criminisi et al., 2012), while MLP provides a tighter control for overfitting and copes well with non-linear problems (Collobert and Bengio, 2004).

The emotion arc feature set (EmoArc) is used for classification in a random forest, multi-layer perceptron, and a convolutional neural network (CNN). For the first two classification methods, I flatten the emotion-segment matrix into an input vector. From these representations, the classifiers can learn which emotion matters for which segment, like, *e.g.*, "high value at position 2", "low value at position 4" and combinations of these characteristics. However, they are challenged by a need

to capture interactions such as "position 2 has a higher value than position 3", or similar relationships at different positions, like "highest value at position around the middle of the book".

To address this shortcoming, I also experiment with a convolution neural network, visualized in Figure 3.2. The upper part of the input matrix corresponds to the emotion-segment matrix from Section 3.4.1. Below, I add k one-hot row vectors each of which encodes the position of one segment. This representation enables the CNN with EmoArc features to capture the development of different emotions between absolute segment positions – it can compare the "intensity" of different emotions over time steps. By considering all emotions through time steps in a text, the CNN can model patterns outside the expressivity of the simpler classifiers.

Formally, the CNN consists of an input layer, one convolutional layer, one max pooling layer, one dense layer, and an output layer. The convolutional layer consists of 32 filters of size $(8 + k) \times 4$. The max pooling layer takes into account regions of size 1×2 of the convolutional layer and feeds the resulting matrices to the fully connected dense layer with 128 neurons.

3.4.3 Meta-Parameter Setting

I choose the following meta-parameters: For RF, I set the number of trees to 250 in BoW and EmoLex and to 430 in

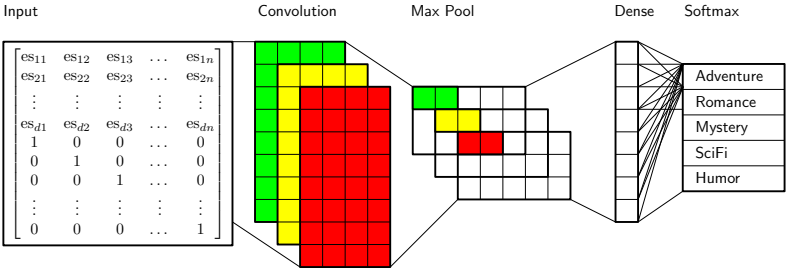


Figure 3.2: Architecture of CNN model.

EmoArc. In MLP, I use two hidden layers with 256 neurons each, with an initial learning rate of 0.01 that is divided by 5 if the validation score does not increase after two consecutive epochs by at least 0.001. Each genre class is represented by one output neuron. For the number of segments in the text, I choose $k = 6$.

3.5 Genre Classification Results

Table 3.2 shows the main results in a 10-fold cross-validation setting. The BoW baseline model shows a very strong performance of 80 % F_1. Limiting the words to those 4,463 which are associated with emotions in EmoLex significantly improves the classification of humorous and science fiction books, which leads to a significant improvement of the micro-average precision, recall, and F_1 by 1 percentage point. This result shows that emotion-associated words predict genre as well as BoW

		Genre														
		adventure			humor			mystery			romance			sci-fi			Micro-Av.		
Model	Features	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RF	BoW	75	89	81✔○	83	51	63✔○	82	78	80✔○	73	74	74✔○	88	87	87✔○	80✔○	80✔○	80✔○
MLP	BoW	73	75	74	69	58	63	70	70	70	66	71	68	85	84	85	74	74	74
RF	EmoLex	69	88	78	91	39	54	81	74	78	75	73	74	83	84	84	77	77	77
MLP	EmoLex	80	79	80✔✖✱	78	78	78✔✖	80	79	79✔✖✱	71	76	73✔✖✱	91	89	90✔✖✱	81✔✖✱	81✔✖✱	81✔✖✱
RF	EmoArc	51	72	59	70	27	39	55	33	41	59	63	61	70	73	71	58	58	58
MLP	EmoArc	55	60	57	56	36	44	49	47	48	57	71	63	72	65	68	58	58	58
CNN	EmoArc	56	60	58○✖	56	43	48○✖	49	46	48○✖	57	70	63○✖	74	68	71○✖	59○✖	59○✖	59○✖
SVM	Ensemble	80	86	83✱	80	78	79	87	79	83✱	79	78	78✱	90	91	91✱	84✱	84✱	84✱

Table 3.2: Results for genre classification on the Gutenberg corpus (percentages). I use bootstrap resampling (Efron, 1979) to test for significance of differences ($\alpha = 0.05$) (a), pairwise among the best models for each feature set (RF BoW, MLP EmoLex, CNN EmoArc) and (b), between the best individual model (MLP EmoLex) and the SVM Ensemble model. Legend: ✔ MLP EmoLex vs. RF BoW, ✖ CNN EmoArc vs. MLP EmoLex, ○ CNN EmoArc vs. RF BoW, ✱ Ensemble SVM vs. MLP EmoLex.

model even though fewer words, and particularly less content-related words are considered. This aspect is further discussed in the model analysis in Section 3.6.3 and Table 3.7. I test for significance of differences ($\alpha = 0.05$) using bootstrap resampling (Efron, 1979), see the caption of Table 3.2 for details.

Among the EmoArc models, I find the best performance (59 % F_1) for the CNN architecture underlining the importance of the model to capture emotion *developments* rather than just high or low emotion values. The EmoArc models significantly underperform the lexical approaches. At the same time, their results are still substantially better than, *e.g.*, a most frequent class baseline (which results in 12 % F_1). Thus, this result shows the general promise of using emotion arcs for genre classification, even though the non-lexicalized emotion arcs represent an impoverished signal compared to the lexicalized BoW and EmoLex models.

This raises the question of whether a model combination could potentially improve the overall result. Table 3.3 quantifies the complementarity of the models: Its diagonal shows true positive counts for each model. The other cells are true positive hits for the column models which were *not* correctly classified by the row model. Therefore, the additional contribution, *e.g.*, by MLP EmoArc over MLP EmoLex consists in 123 additional correctly classified texts. Conversely, 586 texts are correctly classified by MLP EmoLex, but not by MLP

Model	Features	RF BoW	MLP BoW	RF EmoLex	MLP EmoLex	RF EmoArc	MLP EmoArc	CNN EmoArc
RF	BoW	**1616**	110	38	184	73	98	103
MLP	BoW	228	**1498**	215	298	172	182	176
RF	EmoLex	99	158	**1555**	240	72	111	114
MLP	EmoLex	161	157	156	**1639**	133	123	131
RF	EmoArc	503	484	441	586	**1186**	194	197
MLP	EmoArc	536	502	488	584	202	**1178**	100
CNN	EmoArc	520	475	470	571	184	79	**1199**

Table 3.3: Model comparison. Numbers on the diagonal show the numbers of overall true positives for the respective model. Numbers in other cells denote the number of instances correctly classified by the column model, but not by the row model.

EmoArc.

These numbers indicate that my models and feature sets are complementary enough to warrant an ensemble approach. This is bolstered by an experiment with an oracle ensemble. This oracle ensemble takes a set of classifiers and considers a classification prediction to be correct if at least one classifier makes a correct prediction. It measures the upper bound of performance that could be achieved by a perfect combina-

tion strategy. Taking into account predictions from all the models in Table 3.2 yields a promising result of 94 % F_1 (precision=recall=94 %), an improvement of 14 percentage points in F_1 over the previous best model.

Following this idea of a combination strategy, I implement an ensemble model that is an L1-regularized L2-loss support vector classification model that takes predictions for each book from all the models as input and performs the classification via a 10-fold cross-validation. The results for this experiment are given in Table 3.2 on page 75 in the last row. Overall, I observe a significant improvement over the best single model, the MLP EmoLex model.

As the results show, the outcome of my ensemble experiment is still far from the upper bound achieved by the oracle ensemble. At the same time, even the small, but significant, improvement over the best single model provides a convincing evidence that further improvement of the classification is possible. However, finding a more effective practical combination strategy presents a multiaspect problem with vast solution space which I leave for future work.

I now proceed to obtaining a better understanding of the relationship between emotion development and genres.

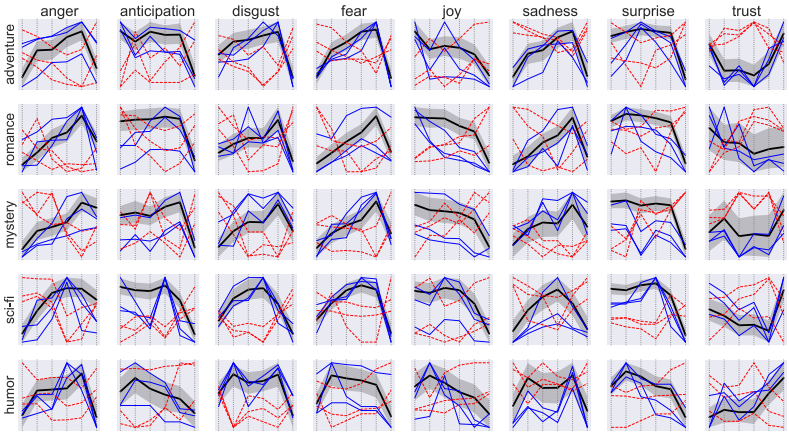


Figure 3.3: Emotion developments per genre. Thick black line: prototypical development. Grey band: 95% confidence interval. Blue lines: 3 most correlated books within each emotion-genre pair. Red dashed lines: 3 least correlated books within each emotion-genre pair.

3.6 Model and Data Analysis

3.6.1 Uniformity of Prototypical Arcs

The results presented in the previous section constitute a mixed bag: even though overall results for the use of emotion-related features are encouraging, the specific EmoArc model was not competitive. I now investigate possible reasons.

My first focus is the fundamental assumption underlying the EmoArc model, namely that *all works of one genre develop*

relatively uniformly with respect to the presence of individual emotions over the course of the plot. I further concretize this notion of *uniformity* as correlation with the *prototypical emotion development for a genre* which I compute as the average vector of all emotion scores (*cf.* Section 3.4.1) for the genre in question.

I formalize the *uniformity* of an emotion arc of a text with scores $\langle \mathrm{es}_1, \ldots, \mathrm{es}_k \rangle$ as the Spearman rank correlation coefficient with the prototypical vector $\langle \overline{\mathrm{es}}_1, \ldots, \overline{\mathrm{es}}_k \rangle$, which is an average of all books in the category. Spearman coefficients range between -1 and 1, with -1 indicating a perfect inverse correlation, 0 no correlation, and 1 perfect correlation. In contrast to, *e.g.*, a Euclidean distance, this measures the emotion arc in a similar manner to the CNN.

Figure 3.3 on page 79 shows the results in an emotion-genre matrix. Each cell presents the emotion scores for the six segments, shown as vertical dotted lines. The thick black line is the prototypical development, and the grey band around it a 95% confidence interval. We see the three most correlated (*i.e.*, most prototypical) books in blue, and the curves for the three least correlated (*i.e.*, most idiosyncratic) books in dashed red.

The figure shows that there are considerable differences between emotions-genre pairs: some of them have narrow confidence bands (*i.e.*, more uniform behavior), such as *fear*, while others have broad confidence bands (*i.e.*, less uniform behav-

	Genre				
Emotion	Adv.	Humor	Myst.	Rom.	Sci-fi
Anger	0.21	0.20	**0.25**	0.28	0.18
Anticipation	0.12	0.10	0.17	0.15	0.16
Disgust	0.17	0.22	0.14	0.21	0.14
Fear	**0.28**	**0.22**	0.19	**0.32**	**0.19**
Joy	0.15	0.09	0.14	0.19	0.16
Sadness	0.21	0.18	0.12	0.25	0.15
Surprise	0.17	0.16	0.19	0.23	0.17
Trust	0.16	0.17	0.07	0.07	0.13

Table 3.4: Average uniformity of emotion-genre pairs measured by Spearman correlation. Highest uniformity per genre marked in bold.

ior), such as *trust* and *anticipation*. Table 3.4 on page 81, which lists the average uniformity (Spearman correlation) for each genre-emotion pair, confirms this visual impression: the emotions that behave most consistently within genres are *fear* (most uniform for four genres) and *anger* (most uniform for *mystery*). In contrast, the emotions *anticipation* and *trust* behave nonuniformly, showing hardly any correlation with prototypical development.

These findings appear plausible: *fear* and *anger* are arguably more salient plot devices in fiction than *anticipation* and *trust*. More surprisingly, *happiness/joy* is not among the most uniform emotions either. In this respect, my findings do not match the results of Reagan et al. (2016): according to my results, *joy*

is not a particularly good emotion to base a genre classification on. I discuss reasons for this discrepancy below in Section 3.7.

At the level of individual books, Figure 3.3 on page 79 indicates that there are "outlier" books (shown in dashed red) with a development that is almost completely inverse compared to the prototype for essentially *all* emotion-genre pairs, even the most uniform ones. This finding can have two interpretations: either it indicates unwarranted variance in my analysis method (*i.e.*, the assignment of emotions to text segments is more noisy than I would like it to be), or it indicates that the correlation between the emotion plot development and the genre is weaker than I initially hypothesized.

As a starting point for a close reading investigation of these hypotheses, Table 3.5 on page 90 lists the three most and least prototypical books for each genre, where I averaged the books' prototypicality across emotions. Note that the list of the least prototypical books contains some well-known titles, such as *La dame aux Camilias*, while the top list contains lesser known titles. A cursory examination of the emotion arcs for these works indicates that the arcs make sense. Thus, it seems that more outstanding literary works literally "stand out" in terms of their emotion developments: their authors seem to write more creatively with respect to the expectations of the respective genres.

However, other factors such as book length and publica-

tion date could also result in a book deviating from the mean. Therefore, the notion of a prototypical arc requires further examination in future work.

3.6.2 Emotion Arcs and Genre Classification

Above, I have established that arcs for some emotions are more uniform than others, and that there are outlier texts for every emotion and genre. But does the degree of uniformity matter for classification? To assess this question, I analyze the average prototypicality among books that were classified correctly and incorrectly for each classification model from Section 3.4.2.

The results in Table 3.6 on page 91 show that the average prototypicality is always higher for correctly than for incorrectly classified books. That being said, there appears to be a relationship between the feature set used and the size of this effect, Δ. This size is smallest for the EmoLex models, higher for the BoW models. It is considerably higher for the EmoArc models (with close values for MLP and CNN models).

I draw three conclusions from this analysis: (1), EmoArc features and models based on them are meaningful for the task of literary genre classification, as evidenced by higher correlation coefficients in the correctly predicted instances. (2), since emotion arcs are exactly the type of information that the CNN EmoArc model bases its classification decision on, emotion

uniformity is indeed a prerequisite for successful classification by EmoArc, and its lack for some genres and emotions explains why EmoArc does not do as well as the more robust BoW and EmoLex models. (3), the difference in correlation ranks between correct and incorrect predictions validates the idea of an ensemble classification scheme and may serve as a starting point for deeper investigation of differences between models in future work.

However, the results presented here should be taken with a grain of salt. Although the idea of genre classification based on emotional arcs is a plausible one, there are certain limitations. As analysis shows, genres are not always uniform with respect to their prototypical development. Hence, accounting for outliers is an issue that one shall consider before the classification. At the same time, a further investigation into the differences between genres with respect to emotional arcs is needed. As Figure 3.3 on page 79 shows the main differences (visually detectable by human eyes) between genres are confidence bands, rather than curves that show trends of limited diversity. This could mean that the differences between genres are subtler than the arc shape and need further analysis.

3.6.3 Feature Analysis of Lexical Models

After having considered EmoArc in detail, I now complete my analysis in this topic by a more in-depth look at the feature

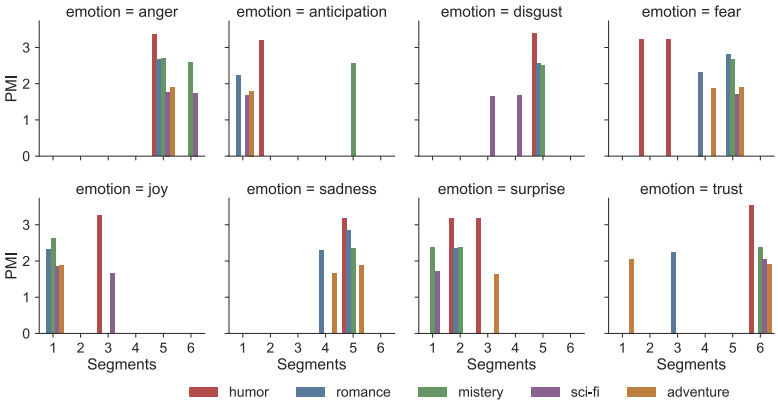


Figure 3.4: Top EmoArc features for each genre ranked according to their PMI values.

level. I focus on features that are most strongly associated with the genres, using a standard association measure, pointwise mutual information (PMI), which is considered to be a sensible approximation of the most influential features within a model.

Table 3.7 on page 92 shows that most strongly associated features with each genre differ in their linguistic status between BoW and EmoLex. For example, for the genre *romance*, most BoW features are infrequent words like specific character names which do not generalize to unseen data (*e.g.*, *Gerard, Molly*). The EmoLex features consist of words related to emotions (*e.g.*, *mamma, marry, loving*). In *mystery*, the most important BoW features express typical protagonists

of crime stories (*e.g.*, *coroner, detective, inspector, Scotland*). For EmoLex, we see similar results with a stronger focus on affect-related roles (*e.g.*, *murderer, jury, attorney, robbery, police, crime*). In sum, we observe that the feature sets pick up similar information, but from different perspectives: the BoW set focusing more on the objective ("what") and the EmoLex set more on the subjective ("how") level.

As a combination of the analysis in Section 3.6.2 with the PMI approach, Figure 3.4 on page 85 visualizes the EmoArc features as "peak" features that fire when an emotion is maximal in one specific segment (*cf.* Section 3.5). The results correspond well to the prominent maxima of emotion arcs shown in Figure 3.3. For the genre of adventure, *e.g.*, *trust* and *anticipation* peak at the beginning. *Sadness, anger*, and *fear* peak towards the end, however, the very end sees a kind of "resolution" with trust becoming the dominating emotion again. At the same time, *anger* and *sadness* seem to be dominating all genres towards the end, and *joy* plays an important role in the first half of the books for most genres.

3.7 Discussion and Conclusion

In this section, I tried to answer Research Question 1: "Does emotion classification performed with dictionaries allow for the classification of genres?" To that end, I analyzed the relation-

ship between emotion information obtained with NRC dictionary and genre categorization. I considered three feature sets corresponding to three levels of abstraction (lexical, lexical limited to emotion-bearing words, emotion arc) and found interesting results: classification based on emotion-words performs *on par* with traditional genre feature sets that are based on rich, open-vocabulary lexical information. My first conclusion is therefore that *emotion carries the information that is highly relevant for distinguishing genres.*

A further aggregation of emotion information into emotion arcs currently underperforms compared to the lexical methods, indicating that relevant information gets lost in the current representation. I need to perform further research regarding this representation as well as the combination of different feature sets, since these appear to contribute complementary aspects to the analysis of genres, as the excellent performance of an oracle shows. The ensemble approach significantly outperforms the best single model but still outperforms the oracle result.

My subsequent, more qualitative analysis of the uniformity of emotion arcs within genres indicated that some, but not all, emotions develop moderately uniformly over the course of books within genres: *Fear* is the most uniform in all genres except mystery stories, where *anger* is more stable. Unexpectedly, *joy* is only of mediocre stability. At the same time, my study of outliers indicates that this conforming to the proto-

typical emotion development of a given genre appears to be a *sufficient, but not necessary* condition for membership in a genre: I found books with idiosyncratic emotion arcs that were still unequivocally instances of the respective genres. As with many stylistic properties, expectations about emotion development can evidently be overridden by a literary vision.

This raises the question of what concept of genre it is that my models are capturing. Compared to more theoretically grounded concepts of genre in theoretical literary studies, my corpus-based grounding of genres is shaped by the books I sampled from Project Gutenberg. Many of these are arguably relatively unremarkable works that exploit the expectations of the genres rather than seminal works trying to redefine them. The influence of corpus choice on my analysis take may also explain the apparent contradictions between my by-emotion results and the ones reported by Reagan et al. (2016)[3], who identified *happiness/joy* as the most important emotion, while this emotion came out as relatively uninteresting in my analysis. My observations about the influence of individual artistic decisions have, however, made me generally somewhat hesitant regarding Reagan et al.'s claim about "universally applicable plot structures".

The results presented in this section suggest that *emotion*

[3] 1,372 books from Project Gutenberg crawled with different filters than used by me.

dictionaries can be used for emotion classification, as evidenced by acceptable genre classification results. More importantly, we have seen that emotion "is in the text". What we do not know yet is where exactly in the text it is. Given the fact, that emotional arcs are built using aggregated emotion information from text, it is not possible to tell if emotions in text belong to characters, to an author, or are not even emotions but rather incidental occurrences of words that have an associated emotion in a dictionary. In other words, what structure do emotions have on a linguistic level? Do they even have a structure? If so, can we automatically detect this structure? I address these questions in detail in the next section.

Genre	Most prototypical	Least prototypical
adventure	*Bert Wilson in the Rockies*, Duffield, J. W.	*Chasing the Sun*, Ballantyne, R. M.
	The Outdoor Girls of Deepdale, Hope, L.	*The Bronze Bell*, Vance, L.J.
	Blown to Bits, Ballantyne, R. M.	*Chester Rand*, Alger, H.
romance	*The Girl in the Mirror*, Jordan, Elizabeth Garver	*La Dame aux Camilias*, Dumas, A.
	The Unspeakable Perk, Adams, Samuel Hopkins	*Through stained glass*, Chamberlain, George
	The Maid of Maiden Lane, Barr, Amelia	*Daddy-Long-Legs*, Webster, Jean
mystery	*The Woman from Outside*, Footner, H.	*The Grell Mystery*, Froest, F.
	The Old Stone House and Other Stories, Green, A.K.	*My Strangest Case*, Boothby, G.
	In Friendship's Guise, Graydon, W.M.	*The Treasure-Train*, Reeve, A. B.
scifi	*The Great Drought*, Meek, S. P.	*Looking Backward, 2000 to 1887*, Bellamy, E.
	The Finding of Haldgren, Diffin, Charles	*Let 'Em Breathe Space!*, Del Rey, L.
	The Tree of Life, Moore, C. L.	*The Second Deluge*, Serviss, Garrett P.
humor	*Captains All and Others*, Jacobs, W.	*Just William*, Crompton, Richmal
	The Rubáiyát of a Bachelor, Rowland, H.	*Baby Mine*, Mayo, Margaret
	The Temptation of Samuel Burge, Jacobs, W. W.	*Torchy and Vee* , Ford, Sewell

Table 3.5: Most and least prototypical books regarding overall emotion development in each genre

Model Family	Classification	Avg. Spearman on +	Avg. Spearman on −	Δ between + and −
BoW	RF	0.185	0.164	0.021
	MLP	0.184	0.170	0.014
EmoLex	RF	0.182	0.176	0.006
	MLP	0.181	0.179	0.002
EmoArc	RF	0.193	0.162	0.031
	MLP	0.206	0.144	0.062
	CNN	0.205	0.145	0.060

Table 3.6: Average prototypicality (measured as correlation with prototypical emotion arc) for books that are correctly (+) and incorrectly (−) predicted by each model. Positive Δ means higher prototypicality for correct classifications.

BoW					EmoLex				
Adv.	Humor	Mystery	Romance	SciFi	Adv.	Humor	Mystery	Romance	SciFi
tarzan	ses	coroner	gerard	planet	hermit	wot	murderer	sally	projectile
damon	iv	kennedy	molly	solar	hut	wan	jury	mamma	rocket
canoes	sponge	detective	willoughby	planets	fort	comrade	attorney	marry	beam
blacks	ay	inspector	fanny	projectile	lion	rat	robbery	tenderness	scientist
indians	says	detectives	clara	mars	tribe	bye	police	loving	blast
ned	wot	trent	maggie	rocket	spear	beer	crime	charity	bomb
savages	wan	scotland	eleanor	rip	jungle	idiot	criminal	love	emergency
spain	mole	murderer	cynthia	jason	swim	jest	murder	marriage	system
whale	ha	rick	yo	phone	rifle	school	suicide	passionate	center
eric	ma	scotty	jill	globe	don	mule	clue	holiday	pilot

Table 3.7: Top ten EmoLex and BoW features by pointwise mutual information values with each genre.

4 Linguistic Structure of Emotion

4.1 Introduction[1]

In Section 3, I provided an analysis of an interaction of emotions and literary genres using a classification approach as a probing tool for understanding if emotion dictionaries are meaningful in the analysis of a literary text.

I find that dictionaries do provide meaningful information, which is evidenced by the results of genre classification based on emotional arcs. However, given the fact, that emotional arcs are built using aggregated emotion information from text, it is not possible to answer the following question: 1) who is the experiencer of an emotion? (emotion does not exist without an experiencer), and 2) what causes an emotion? Therefore, in this chapter, I will present the results of my work on modelling emotion structure in literary text. Given the task difficulty, I will primarily focus on answering the following question:

[1]This chapter is an extension of Kim and Klinger (2018).

Research Questions

Which parts of the emotion structure are the most difficult for the classification task and are the extraction tasks of different parts of emotion structure interrelated?

To that end, I will set up an annotation task and relevant modelling experiments, which I describe in detail in the subsequent sections.

4.2 Contributions

My main contributions related to this topic are the following:

1. I present the first resource of fictional texts annotated with semantic roles of emotions, experiencers, causes, and targets.

2. I show that emotion annotation that takes into account not only strong emotion indicators ("afraid"), but also implicit emotions ("shaking fingers") is valuable for the study of the language of emotions.

3. I provide results of baseline models to predict emotion words and roles separately

4. I show that the prediction performance of all subtasks benefits from joint prediction of experiencer, emotion words, and targets.

4.3 Annotation Task

The annotation and modelling tasks I describe in the following sections can broadly be situated alongside existing research on semantic role labeling (Palmer et al., 2010). In particular, this work to certain extent follows the concept of *directed emotion*, as defined in FrameNet (Baker et al., 1998; Fillmore et al., 2003), and extends the work of Ghazi et al. (2015), who focus on detecting emotion stimulus in the FrameNet exemplary sentences using conditional random fields. I do not use FrameNet annotations in my thesis, as there is not sufficient annotations of fictional text and most annotations do not span more than a single sentence (which may be insufficient for my purpose).

The goal of my annotation project is to create a dataset of excerpts from fictional texts that are annotated for the phrases that lead to the association of the text with an emotion, the experiencer of the emotion (a character in the text, if mentioned), the target and the cause of the emotion, if mentioned (*e.g.*, an entity, or event). An example of such an annotation is shown in Figures 4.1 and 4.2 on page 96. As it can be seen from these depictions, each annotation includes textual span

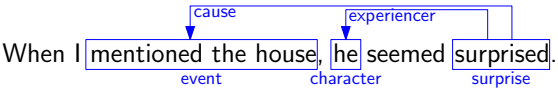


Figure 4.1: Example annotation from Hugo (1885), with one character, an emotion word, and event and cause and experiencer annotations.

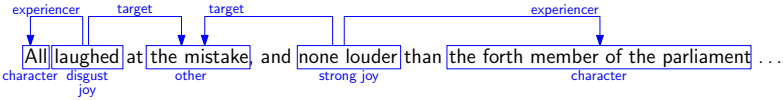


Figure 4.2: Example annotation from Stimson (1943), with two characters who are experiencers of different emotions. Disgust and joy are annotated as a mixture of emotions. Both emotions have the same target.

annotations such as emotions, characters, events, as well as relation annotations that establish relations between different text spans (cause, experiencer, target). In the following, I describe the conceptual background for each annotation layer in detail. The complete annotation guidelines are available online together with the corpus[2].

[2] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reman/

4.3.1 Phrase Annotation

Emotion

We conceptualize emotions as one's experience that falls in the categories in Plutchik's classification of emotions, namely *anger*, *fear*, *trust*, *disgust*, *joy*, *sadness*, *surprise*, and *anticipation*. In addition, we allow annotation with the class *other emotion* that covers cases when the emotion expressed in the text cannot be reliably categorized into one of the predefined eight classes. A list of the emotions along with example realizations can be found in Table 4.6 on page 126.

Annotators are instructed to prefer span annotations of key words (*e.g.*, "afraid"), except cases when emotions are only expressed with a phrase (*e.g.*, "tense and frightened") or indirectly (*e.g.*, "the corners of her mouth went down"). Additionally, emotion spans are marked to be intensified (*i.e.*, amplified), diminished (*i.e.*, downtoned) and negated without marking the modifier or including the modifier. Each span is associated with one or more emotions (exemplified in Figure 4.2 on page 96).

Entity

We conceptualize entities as mentions of something that has a clear identity of a person, object, concept, state, or event (see Table 4.7 on page 127). Entities are only annotated if they are

experiencer, cause, or target of an emotion.

Character An entity that acts as a character in the text. Character annotation should not omit important information (*e.g.*, the annotation of "the man with two rings of the Royal Naval Reserve on his sleeve" is preferred over only annotating "the man").

Event An event is an occasion or happening that plays a role in the text. Events can be expressed in many ways (see Table 4.7 on page 127) and annotators are instructed to label the entire phrases including complementizers or determiners.

Other This is an umbrella concept for everything else that is neither a character nor an event, but fills as relation, described in the following.

4.3.2 Relation Annotation

Relations are semantic links between an emotion and other text spans and can be of type *experiencer*, *cause*, and *target*. In addition, we partially annotate *coreferences* to link personal pronouns to proper nouns. All relations, except Coreference, can only originate from the emotion annotations.

Experiencer The experiencer relation links an emotion span and entity of type *character* who experiences the emotion. If the text contains multiple emotions with multiple experiencers, they all are subject to relation annotation.

Target The target relation links an emotion span and entity of any type towards which the emotion experienced by the experiencer is directed. If there are multiple targets of the emotion, then all of them should also be included in the relation annotation. See Figure 4.2 (page 96) for the example of a target annotation.

Cause The cause relation links an emotion span and entity of any type, which serves as a stimulus, something that evokes the emotion response in the experiencer. If there are multiple causes for the emotion, then all of them are included in separate relation annotations.

Coreference The annotators are instructed to annotate as an experiencer the character that is the closest to the emotion phrase in terms of token distance. If the closest mention of the character is a pronoun and the text provides a referent that has a higher level of specificity than the pronoun (*i.e.*, a proper noun or a noun denoting a group or class of objects), then the annotators are asked to resolve the coreference.

4.4 Corpus Construction and Annotation

4.4.1 Selection

The corpus of 200 books is sampled from Project Gutenberg[3]. All books belong to the genre of fiction and were written by authors born after the year 1800.

I sample consecutive triples of sentences from this subsample of books. A triple is accepted for inclusion for annotation if the middle sentence includes a word from the NRC dictionary (Mohammad and Turney, 2013). I consider this middle sentence the target sentence and the annotators are instructed to label emotions in this second sentence only. Experiencers, causes and targets are annotated in the whole sentence triple if they refer to an emotion in the target sentence.

The sampling procedure is motivated by the observation that triples of sentences sampled with emotion dictionary show the best coverage in terms of the roles that are associated with the emotion. Ghazi et al. (2015) annotate only one sentence and speculate whether adding one sentence before and after will lead to better results. To check their hypothesis, I conduct a small pre-study experiment by extracting 100 random sentences from the Project Gutenberg with the NRC dictionary and analyze how often the roles of experiencer, cause, and target are found in the target sentence and in the window

[3] `http://www.gutenberg.org/`

of up to five sentences before and after. The analysis shows that 98% of the texts include the experiencer in the target sentence, while cause and target is found in the target sentence in 67% of the texts. Another 29% of the texts include cause and target in the window of one sentence before and after the target sentence. The remaining texts include cause and target in the window of two (2%), three (1%), and four (1%) sentences around the target sentence. I therefore opt for three-sentence spans as they provide enough information regarding "who feels what and why" without creating unnecessary annotation overhead (cumulatively, 96 % of cause and target are found in such sentence triples).

4.4.2 Genre and Author Composition

Table 4.1 on page 119 shows 10 most frequent categories of fiction represented in the corpus along with the most frequently occurring authors in each category.

4.4.3 Annotation Procedure

The annotations were generated in a multistep process, visualized in Figure 4.3 on page 120. The people involved in the annotation were either *annotators* or *experts*, whose roles did not overlap. The annotations (of spans and relations) were performed by three graduate students of computational linguistics (two native English speakers, one non-native speaker)

within a three-month period. Arising questions were discussed in weekly meetings with the experts (the authors of the paper) and the results documented in the annotation guidelines. I used WebAnno[4] Yimam et al. (2013) as annotation framework. In the following, I discuss the four steps of generating the corpus.

Step 1: Emotion phrase annotation The *annotators* first decide whether the text expresses an emotion and which emotion it is. If any exists, they label the phrase, which led to their decision. The annotators were instructed to search for emotions that are expressed either as single words or phrases.

Step 2: Emotion phrase aggregation In the previous step, annotators generate set of annotations. In this step, the *expert* heuristically aggregates all spans that overlap between annotators in a semi-automatic process: Concrete emotions are preferred over the "other-emotion", annotations with modifier are preferred over annotations without, and shorter spans are preferred over longer spans. Overlapping annotations with different emotion labels are all accepted.

Step 3: Relation annotation *Annotators* are given the same texts they annotated for emotions in Step 1 with the aggregation from Step 2 from all annotators. Therefore, all

[4] `https://webanno.github.io/webanno/`

annotators see the same texts and annotations as input in this step. For each emotion, the task is to annotate entities that are experiencers, targets, or causes of the emotion and establish relations between them. The annotators were instructed to tag only those entities that have a role of an experiencer, cause, and target. The decision on the entity and relation annotation is made simultaneously: For each emotion the annotators find who experiences the emotion (which *character*) and why (because of event, object, or other character).

Step 4: Relation annotation aggregation and adjudication

This final step is a manual *expert* step: Aggregate the relation annotations provided by the annotators. Heuristically, shorter spans for entities are preffered without discarding common sense. For instance, consider the phrase "[...] *wishing rather to amuse and flatter himself by merely inspiring her with passion*". "*Wishing*" is labelled as emotion. One annotator tagged "*to amuse and flatter himself by merely inspiring her with passion*" as event, another tagged only "*by merely inspiring her with passion*", which is incomplete, as the target of the emotion is the act of amusing and flattering oneself. The purpose of having an *expert* annotator is, thus, to perform an annotation aggregation from multiple annotators in cases where automatic aggregation is not possible.

All annotations of all annotators are published.

4.5 Results

In the following, I first discuss annotation statistics and then provide results of the models trained on the resource.

4.5.1 Inter-annotator Agreement and Consistency of the Annotations

I use pairwise Cohen's Kappa coefficient (κ) on the token level and F_1 on a phrase level with exact and fuzzy match to calculate the agreement of the phrase annotation and F_1 to estimate the agreement of the relation annotation. For F_1 calculation, I use two approaches: *strict* that requires labels and spans to be identical, and *fuzzy* that accepts an annotation to be a true positive if the annotations of two annotators overlap by at least one token. Table 4.2 on page 121 reports the agreement scores for emotion, entity, and relation annotations between each pair of annotators.

Joy has the highest number of instances (336) and the highest agreement scores (average $\kappa = 35$), followed by *fear* ($\kappa = 0.30$) and *sadness* ($\kappa = 0.24$). *Other emotion* has the lowest agreement with average $\kappa = 0.07$. For entity annotation, especially for *character* annotation, the agreement is higher, with the highest agreement between two annotators being $\kappa = 0.63$.

The agreement on the *event* and *other* entities is low ($\kappa = 0.23$ and 0.14 and $F_1 = 25$ and 14, respectively). This is presumably the case because event annotations are often comparably long. This also holds, to a lower extent, for *character* annotations. If partial overlaps to count as a match are allowed, the average F_1 increases to 57 for *character* (an increase of 4 percentage points (pp)), 44 for *event* (increase by 19 pp), and 23 for *other* category (increase by 9 pp).

For relation annotations, higher agreement scores are also observed with fuzzy evaluation (F_1 increase for *experiencer*, *cause* and *target* by 10 pp, 7 pp, and 12 pp respectively). These results are in line with previous studies on emotion cause annotation Russo et al. (2011), and show that disagreements mainly come from the different spans of the entities, though they overlap.

4.5.2 Difficulties with Obtaining High Agreement

As I showed in Section 4.5.1, the agreement across all annotation layers is comparably low. There are several reasons for that. The cause and target of the emotion are not always clearly recognizable in the text and are also subjective categories (two annotators may find two different causes for the same emotion), hence the low agreement scores across all categories. The only exception are *experiencer* category, which is the most reliable among all annotations and match the sub-

stantial agreement scores of character annotation (the only type of entities that can be involved in an experiencer relation).

I illustrate the difficulties the annotators face when annotating emotions with roles with the following example:

*"**They** had never seen ... what was really **hateful** in his face; ... they could only express it by saying that the arched brows and the long emphatic chin gave it always a look of being lit from below ..."*

All annotators agree on the character ("they") and the emotion ("hateful" expressing disgust). Similarly, both annotators agree that the disgust is related to properties of the face which is described, however, one annotator marks "his face" as target, the other marks the more specific but longer "the arched brows and the long emphatic chin gave it always a look of being lit from below" as cause.

If we abstract away from the text spans, both annotators agree that the emotion of disgust has something to do with "his face", however they disagree on the target annotation and the cause annotation. So, though conceptually, the annotations by two people are similar, this is not reflected in the agreement scores.

4.5.3 Corpus Details

Tables 4.3 on page 122 and 4.4 on page 123 show the total number of annotations for each category. The REMAN corpus consists of 1720 sentence triples, 1115 of which include an emotion. For the emotion category, *joy* has the highest number of annotations, while *anger* has the lowest number of annotations. In most cases, emotion phrases are single tokens (*e.g.*, "monster", "irksome"), out of which 47% on average are found in the NRC dictionary. *Other emotion* has the largest proportion of annotations that span more than one token (36% out of all annotations in this category), which is in line with my expectation that lower levels of specificity for emotion annotation make it more difficult to find a single token that indicates an emotion.

For entities, *character* has the highest number of annotations. As one can see, the *experiencer* relation dominates the dataset (48%), followed by *target* (28%) and *cause* (24%) relations. Note that each character can experience more than one emotion, hence the difference between the number of characters and the experiencers. Table 4.4 (page 123) also shows how many times each emotion triggered certain relation. In this sense, *joy* has triggered the most *experiencer* and *cause* relations, which is still related to the prevalence of the annotations for this emotion in the dataset.

4.6 Models

I provide models for automatically predicting the annotated structures. The goal is to 1) have a simple but transparent model, and 2) to show that the data is not meaningless, which can be the case due to low IAA scores.

To predict parts of emotion structure, I map the relations to span prediction tasks. This is feasible because characters, entities, and other were only annotated if they fill one of the roles, *experiencer*, *target*, or *cause*. Therefore, the prediction task boils down to a sequence prediction task of emotion phrases (for the different emotions) and the potential mentions of experiencers, targets, and causes. Note that the actual relation information in this simplification is lost.

Consider an example depicted in Figure 4.1 on page 96: The phrase "I mentioned the house" is labelled as an event and is assigned a role of a *cause* for the emotion of *surprise*, and the word "he" is labelled as a character and is assigned a role of an *experiencer* of the same emotion. I represent these relationships by tagging "I mentioned the house" as *cause* and "he" as *experiencer* using inside-outside-beginning (IOB) encoding capturing the text spans that are linked by relations with an emotion.

I use two sequence labelling models, conditional random fields (CRF) (Lafferty et al., 2001) and bidirectional long short-term memory networks with a CRF layer (biLSTM-CRF). Both

models provide a good performance in sequence prediction tasks (Benikova et al., 2014; Huang et al., 2015). In addition, I analyze the difficulty of predicting the emotion for a full sentence triple, independent of segments. In the following, I further specify the experimental setting in detail.

4.6.1 Experimental Setting

Experiment 1: Coarse-grained emotion classification In this experiment, the task is to classify the emotions which occur in the sentence triple which forms the instance under consideration. This is therefore a coarse abstraction of the structured prediction tasks presented in this paper. However, this constitutes the most straight-forward task in emotion analysis. I use a dictionary-based approach and a bag-of-words-based classifier.

For the dictionary-based classification, I take the intersection between the words in the triple and NRC dictionary and assign the triple with the corresponding emotion labels. The F_1 score is calculated by comparing the set of labels predicted by dictionaries against the set of gold labels for each triple. The gold labels come from the annotation of words and phrases within each triple. For the BOW approach, I convert each triple into a sparse matrix using all words in the corpus as features. I then classify the triples with a multi-layer perceptron with three hidden layers, 128 neurons each, with an initial learning rate

of 0.01 that is divided by 5 if the validation score does not increase after two consecutive epochs by at least 0.001.

Experiment 2: Fine-grained emotion and role detection In this experiment, I evaluate the performance of fine-grained emotion and role (experiencer, target, and cause) prediction in a sequence labelling fashion, as described above. I instantiate separate CRF and biLSTM-CRF models for each relation, as some annotations overlap (*e.g.*, experiencers can also be targets/causes). The CRF uses part-of-speech tags (detected with spaCy[5] (Honnibal, 2013)), the head of the dependency, if it is capitalized, and offset conjunction with the features of previous and succeeding words as features. For the *emotion* category, I use the presence in the NRC dictionary in addition and, for *experiencer*, the presence in a list of English pronouns. The training is done for 500 iterations with L-BFGS (Liu and Nocedal, 1989) and L1 regularization.

The biLSTM-CRF model uses a concatenated output of two biLSTM models (one trained on word embeddings with dimension 300, and one trained on character embeddings from the corpus with dimension 100) as an input to a CRF layer. The word embeddings that I use as input are pre-trained on Wikipedia[6] using *fastText*. The model is parameterized with

[5] `https://spacy.io/`

[6] As available at `https://github.com/facebookresearch/fastText` Bojanowski et al. (2017)

Adam as an optimization, a dropout value of 0.5, and trained for 100 epochs with early stopping if no improvement is observed after ten consecutive epochs.

Experiment 3: Potential for joint modelling of emotion and role prediction The goal of this experiment is to understand if joint modelling of relations has the chance to contribute over learning each relation separately. To that end, I analyze the potential interactions between predictions with gold labels of all other predictions. Specifically, when training the models, I provide the classifier with the information which sequence of tokens is an experiencer (in the case of emotion phrase prediction) and which sequence of tokens is an emotion (in case of experiencer, cause, and target detection).

4.6.2 Results and Discussion

The results of all the experiments are summarized in Table 4.5 (page 124). I evaluate the models in the same way I use F_1 for inter-annotator agreement: Firstly, by accepting a TP if it is exactly found (exact) and secondly, if at least one token is overlapping with the annotation (fuzzy).

Experiment 1 Emotion classification with dictionaries and bag of words show mediocre performance. The recall with the dictionary classification is comparably high ($F_1 = 83$), which is

due to the fact that texts were sampled using these dictionaries. However, as we said earlier, annotators are free to label any words and phrases as emotion-bearing, hence low precision and F_1 score. The MLP with BOW features does not perform better but shows increased precision at the cost of lower recall.

Experiment 2 As results of this experiment show, the recall is low for all categories. A presumable reason is, as discussed in Section 4.5, that substantial number of emotion annotations are words or phrases that are not found in the NRC dictionary. On average, only 46% of emotion annotations are single tokens that can be found in the NRC dictionary, but for some emotions this number is much lower (only 14% of *anticipation* annotation). The same applies to cause and target categories, as in most cases these are long spans of text (*e.g.*, 94% of events are multiword expressions). This explains zero F_1 score for cause prediction with CRF and biLSTM-CRF and a better performance for target prediction with CRF, taking into account that most target relations is triggered by characters, 75% of which are single tokens (see Table 4.4 on page 123).

The highest precision and F_1 across all categories is observed for the *emotion* category with biLSTM-CRF (strict $F_1 = 43$ and fuzzy $F_1 = 48$). The strict F_1 is by 12 pp higher than predicted with dictionaries and with BOW in text classification experiment.

The *experiencer* category is second best, however, the recall for this category is still very low. This can be explained by the fact that experiencers are expressed in the text mostly as personal pronouns. As far as the number of personal pronouns in the texts is relatively low (13% of all tokens in a sentence on average), and only a small fraction of them act as experiencers ($<$ 1% of all tokens in a sentence on average), the classifier cannot learn when an entity is an experiencer or not.

The results of this experiment suggest that prediction of emotion structure is a difficult task hindered by such hard-to-model-variables as emotion *cause* and *target*. Therefore, future work in this direction shall take into account the fact that using relatively simple feature sets and model architectures may not lead to the desired performance. At the same time, one may reconsider the approach to cause and target modelling by predicting, for example, parts of the *constituency tree* corresponding to the emotion cause or target rather than actual textual spans. This, in turn, may result in a deeper linguistic analysis of the emotion realization in text, which can be of interest for some literary scholars.

Experiment 3 The goal of this experiment was to estimate if joint modelling of emotion and roles is feasible. We observe that, for the *emotion* category, F_1 increases by 5 pp in strict and by 9 pp in fuzzy evaluation if the classifier is provided with

the information, which sequence of tokens is an *experiencer*. For *experiencer* prediction, F_1 increases by 20 pp in strict and by 22 pp in fuzzy evaluation if the classifier is told which word or sequence is labelled as emotion. These results indicate the complementarity of both categories. A qualitative study on a subsample of linguistic properties of emotions and experiencers shows that when the emotion expression and experiencer are parts of the same phrase (verb or adjectival phrase), the emotion word serves as a head to the word that represents an experiencer. Hence, the classifier is able to partially learn that any phrase that is a part of the emotion phrase, whose head is a personal pronoun or a proper name, is a potential *experiencer*.

The same applies to *experiencer*: if the head of the governing phrase is an emotion, then the head of the current phrase is a potential *experiencer* (*e.g.*, as shown Figure 4.4 on page 125). However, due to variability of emotion expressions, this cannot always be the case.

4.7 Discussion and Conclusion

In this chapter, I presented a structured approach that was built on top of emotion annotations and machine learning models that take into account sequential information.

The inter-annotator agreement and sequence labelling results of the baseline model show that the task of annotating

emotions and corresponding roles, as well as their subsequent prediction is a difficult one. A high variability of emotion expressions (see Table 4.6 on page 126) and a variability of cause and target expressions make it hard. As I mentioned in the beginning of this chapter, my work can be seen as a subtask of semantic role labelling and future work could make use of existing sophisticated methods existing in this field to tackle these issues. At the same time, the dataset presented in this chapter provides interesting and valuable insights in the language of emotion expression and, therefore, is useful to the community of linguists who are interested in the study of linguistic properties of emotions.

However, developing such a resource has its limitations: Due to the subjective nature of emotions, it is challenging, if not impossible, to come up with an annotation methodology that would lead to less disparate annotations. That is in line with previous research. For instance, Schuff et al. (2017) and Russo et al. (2011) show that the models achieve higher performance when trained on the aggregated-annotation dataset, as opposed to datasets that included only majority vote annotations.

Another difficulty arises from the nature of the texts used. Fictional texts are highly metaphoric and full of allusions and metonymies, which requires thoughtful reading (often reading between the lines) and a broader context. However, this is something that the annotators do not have: all the context

they have at their disposal is a triple of sentences, each of which can rely on information that is available in other parts of the book, but not in the annotation unit. Therefore, it is not always possible to annotate the cause, target, or even the experiencer. This is a trade-off: On the one side, I did not want to annotate full books to have a representative corpus. On the other side, I might not have provided sufficient context. Future work will therefore aim at better understanding how to preselect the relevant context that is needed for reliable annotation and secondly use such knowledge for a follow-up annotation project.

Nonetheless, I am confident that the dataset presented in this section is useful to linguists and digital humanities scholars, as it can be used to analyze interactions of emotions, characters, and events in fictional texts. Researchers studying language and emotion may get valuable insights into the language of emotion expression in general.

The dataset presented here constitutes a difficult task for semantic role labelling. My experiments suggest that the prediction of emotions with their roles is a task that should be tackled with joint models. Therefore, this corpus adds an interesting relation extraction task to the set of existing challenges.

Some of the results presented here suggest that authors use not only single words to describe an emotion. In fact, on average, 26% of all emotions in the corpus are expressed with two

or more words, that is with phrases. Moreover, out of all emotion annotations consisting of one token, only 53% of words, on average, are found in NRC dictionary. This finding suggests that using dictionaries for emotion cues detection is at least not optimal, as emotion dictionaries normally do not include phrases. One way to tackle this problem would be relying on semantic similarity between words in dictionary and words and phrases of the analyzed text (*e.g.*, Agrawal et al. (2018)).

In this chapter, I raised a question "Which parts of the emotion structure are the most difficult for the classification task and are the extraction tasks of different parts of emotion structure inter-related?" I find that such roles as *target* and *cause* are the most difficult for modelling, which is largely explained by the fact that they can take arbitrarily long event descriptions (94% of all event annotations $\geq$ 2 tokens). On the other hand, the roles of *experiencer* and *emotion cue* can be modelled with varied success depending on the model and features. These two roles also seem to be inter-related, as suggested by the joint modeling experiment.

To conclude, I found that annotating emotion is an extremely challenging task. The main challenges in my project were connected to achieving high agreement, both for the emotion in general, and for such parts of emotion structure as emotion *cause* and *target*. I discuss the issues related to this kind of annotation throughout this chapter, and yet, I would like to

rephrase everything that has been said with a single sentence. The main bottleneck in the annotation, as I see it now, is related to the *relative freedom that annotators had when working with text*. Specifically, I did not impose any linguistic constraints on the annotators when working with text, although this could have had a positive impact on the quality of the annotations. For example, providing the annotators with a syntactic tree for each of the sentences could help them to better see the connections between different parts of the sentence. At the same time, it would be possible to include specific instructions in the guidelines for *cause* annotation, *e.g.*, that it should not cross the borders of a constituent. An even more extreme alternative would be to ask the annotators to annotate a head of a phrase as a *cause* or *target* of an emotion. Such annotation would require significant amount of preprocessing and preparatory steps, but could have lead to much higher agreement scores. I recommend that future work on similar annotation topics take these considerations into account.

Subject headings	Most frequent author	# texts
Fiction, Christian fiction	MacDonald George	178
Historical fiction (translations), Epic literature	Hugo Victor	107
Social fiction	Dostoevsky Fyodor	63
Domestic fiction, Single women	Gissing George	45
Young men, Bildungsroman	Thackeray William	42
Love stories	James Henry	38
Didactic fiction	Eliot George	36
Political fiction	Atherton Gertrude Franklin Horn	35
Historical fiction (translations), France	Dumas Alexandre	35
German fiction (translations), Social classes	Freytag Gustav	22

Table 4.1: Most frequent subject headings and authors in the corpus. Subject headings are taken from Project Gutenberg metadata and are shortened for readability.

Figure 4.3: A visualization of the multi-step annotation process.

		a ↔ b			b ↔ c			a ↔ c		
	Type	κ	strict F_1	fuzzy F_1	κ	strict F_1	fuzzy F_1	κ	strict F_1	fuzzy F_1
Emotion	anger	.25	25	39	.15	15	38	.18	18	33
	anticipation	.09	9	23	.07	7	20	.18	18	39
	sadness	.32	32	41	.22	23	41	.19	20	29
	joy	.38	39	50	.40	40	55	.28	28	44
	surprise	.26	26	43	.22	23	33	.27	27	37
	trust	.17	17	26	.14	14	21	.12	13	32
	disgust	.23	23	41	.10	10	26	.19	19	31
	other	.07	7	7	.06	6	11	.08	8	22
	fear	.35	35	48	.28	28	35	.28	28	41
Entity	character	.63	63	68	.48	49	51	.48	48	54
	event	.29	31	60	.09	10	30	.32	34	44
	other	.11	12	28	.11	11	18	.20	21	23
Relation	experiencer		65	73		48	57		46	55
	cause		20	28		34	39		26	32
	target		27	36		18	29		14	28

Table 4.2: Pairwise inter-annotator agreement for the phrase annotation and relation annotation. F_1 is in %. Regarding the relation scores, in strict F_1, a TP holds if the relation annotation is the same and the entity it points to has the same label and span. In fuzzy F_1, a TP holds if the relation annotation is the same and the entity it points to is the same, but the span boundary of the entity is not necessarily the same.

				Modifier			Annotation Length							
	Type	Total	Adjudic.	strong	weak	neg.	1 token		≥ 2 token		in NRC1		in NRC2	
Emotions	anger	192	156	5	12	7	106	68%	50	32%	36	33%	11	22%
	anticipation	248	201	5	3	11	161	80%	40	20%	28	17%	3	8%
	disgust	242	190	2	7	14	144	76%	46	24%	74	51%	16	34%
	fear	254	183	11	16	17	145	79%	38	21%	93	64%	20	52%
	joy	434	336	31	20	28	289	86%	47	14%	184	64%	29	61%
	sadness	307	224	10	2	13	168	75%	56	25%	100	59%	30	53%
	surprise	243	196	12	4	7	156	80%	40	20%	105	67%	19	47%
	trust	264	232	3	3	33	191	82%	41	18%	66	34%	26	63%
	other emotion	432	207	4	4	4	133	64%	41	36%	52	39%	0	0%
Entities	character	2072	1715				1288	75%	427	25%				
	event	858	615				38	6%	577	94%				
	other	771	485				114	24%	371	76%				

Table 4.3: Corpus statistics for emotions annotations. Columns indicate the number of times each emotion was annotated. "in NRC1" shows how many of 1 token annotations are in the NRC dictionary (percentage is given relative to 1 token annotations). "in NRC2" shows how many multi-word annotations include at least one word from NRC.

			Emotion that triggered the relation									Entities involved		
Relation	Total	Adjudicated	anger	anticip.	disgust	fear	joy	other	sadness	surprise	trust	char.	event	other
experiencer	2113	1717 48%	137	164	130	173	309	210	216	171	207	1704		
cause	1261	840 24%	48	45	70	95	174	74	134	125	75	87	398	343
target	1244	1017 28%	106	129	125	96	135	121	62	80	163	444	315	257
overall relations	4618	3574 77%	291	338	325	364	618	405	412	376	445	2238	717	601

Table 4.4: Corpus statistics for relation annotation. Columns indicate the number of times each role was assigned to an entity and how often the respective emotions are in relation to the entity.

Category	Annotations	Exp	Model	Features	Strict P	Strict R	Strict F_1	Fuzzy P	Fuzzy R	Fuzzy F_1
Emotion	1925	1	Rule-based	dictionary	19	83	31			
		1	MLP	BOW	55	21	31			
		2	CRF	all + dictionary	56	6	11	56	6	11
		3	CRF	all + dictionary + experiencer	55	9	16	69	12	20
		2	biLSTM-CRF	embeddings	57	35	43	62	39	48
Experiencer	1717	2	CRF	all + person	50	2	4	50	2	4
		3	CRF	all + person + emotion	74	15	24	78	15	26
		2	biLSTM-CRF	embeddings	49	21	30	49	21	30
Target	1017	3	CRF	all + emotion	50	3	6	50	3	6

Table 4.5: Results in % for the different experiments. F_1 for *cause* with CRF and biLSTM-CRF and for *target* with biLSTM-CRF is zero and therefore not shown here. The column "Exp" refers to the experimental settings described in Section 4.6.1. Note: Exp. 1 has no fuzzy scores as it does not predict spans, but classifies the whole textual instance.

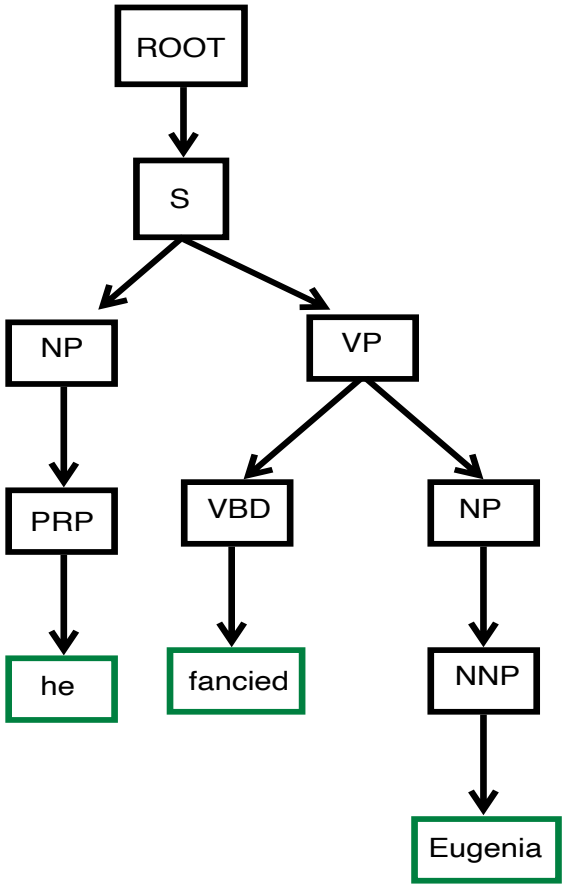


Figure 4.4: Example of an emotion-bearing word ("fancied") governing the experiencer ("he") and target ("Eugenia") of emotion.

	Concept Value	Examples
Emotion	Anger	*angry, defend themselves by force, break your little finger, loss of my temper*
	Anticipation	*want, wish, wholly absorbed, looked listlessly round, wholly absorbed*
	Disgust	*repellent, cheap excitement, turn away from, beg never to hear again*
	Fear	*horrified, tense and frightened, shaking fingers*
	Joy	*cheerful, grateful, boisterous and hilarious, violins moved and touched him*
	Sadness	*failed, despair, the cloudy thoughts, staring at the floor*
	Surprise	*perplexing, suddenly, petrified with astonishment, loss for words, with his mouth open*
	Trust	*honor, true blue, immeasurable patience*
	Other	*careful, brave, had but a tongue, break in her voice, bit deeply into his thumb*
Modifier	strong	*I loved her* ***the more***
	weak	*with a* ***little*** *pity*
	negated	***could not*** *be content*
Entity	character	*the chairman of the board*
	event	*marry a man I did not love, because of his gold*
	other	*Lily's beauty*

Table 4.6: Concepts used for the phrase annotation layer.

Entity type	**Linguistic realiz.**	**Examples**
Character	noun phrase	*his son*
	adjectival phrase	*old man*
Event	verb phrase	*Mrs. Walton had got another baby.*
	adverbial phrase	*Jesus spoke unkindly to his mother when he said that to her.*
	prepositional phrase	[. . .] *giving her up.*
	clause	[. . .] *what she said to him* [. . .]
	noun phrase	*the journey*
Other	adjectival phrase	[. . .] *old age* [. . .]
	noun phrase	[. . .] *the heavens and the earth.*
	tense phrase	*She was the only treasure on the face of the Earth that my heart coveted.*

Table 4.7: Typical linguistic realization of different parts of emotion structure.

5 Emotion-Informed Networks of Characters

5.1 Introduction[1]

In Section 4, I discussed the difficulties associated with prediction of emotion structure. Particularly, we saw that *cause detection* poses the main challenge for the sequence labeling classifiers, as *emotion causes* presented in my dataset are mostly events, and hence, are more difficult to detect. This issue, however, does not directly impede me in my research, as emotion analysis does not boil down to mere understanding of emotion causes. As I mentioned in the opening chapter to this thesis, emotion is at the center of most narrative fiction and there are different aspects to it. One such aspect is the relationship between characters. Character relationship can be thought of as brief or lasting episodes in which one or more characters feel

[1]This chapter is an extension of Kim and Klinger (2019b)

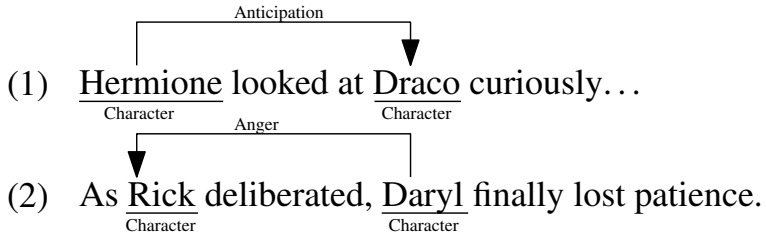


Figure 5.1: Examples of character relationships. The arrow starts at the experiencer and points at the causing character.

certain emotions towards each other[2].

Some studies show that character interactions alone are able to move the plot forward (Booth, 2012, p. 321). This is hardly surprising, as majority of fictional stories are centered around characters in conflict (Ingermanson and Economy, 2009) which interact, grow closer or apart, as each of them has ambitions and concrete goals (Ackerman and Puglisi, 2012, p. 9). These observations suggest that focusing on characters and their relationship may benefit the study of emotions in literary context, as emotions in text are unlikely to exist in isolation from someone who is experiencing them.

Consider Figure 5.1 that depicts two examples of character interactions at the text level. We see that in each example there is one character who feels an emotion directed at another

[2]I use "character interaction" and "character relationship" interchangeably throughout this work.

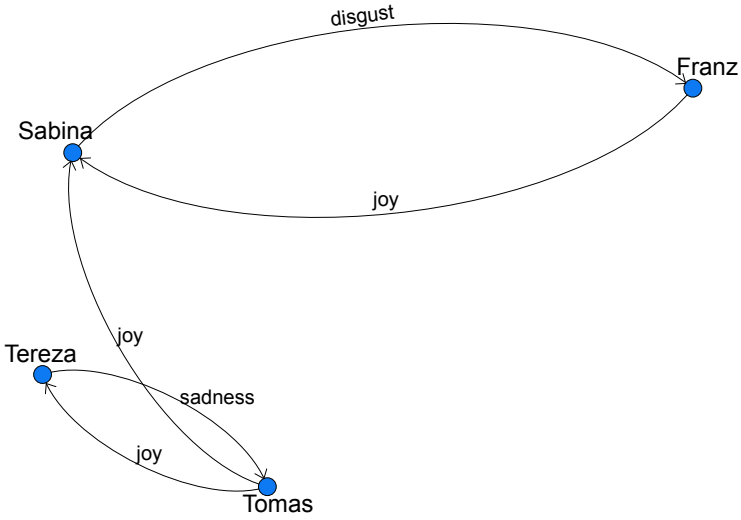


Figure 5.2: An example of character relationship network based on *The Unbearable Lightness of Being* by Milan Kundera.

character. Such interactions may recur periodically throughout a story. Provided that they are thoroughly documented, in the end, one may build a "graph" of character interactions in a story, where each interaction is labelled with an emotion. In other words, it becomes possible to generate a social network of emotion interactions. An example of such a social network is shown on Figure 5.2.

One can see that characters of a story are connected in a directed graph and graph edges are labeled with emotions. Each pair of nodes, thus, has a relationship $(C_{\text{exp}}, e, C_{\text{cause}})$, in which the character C_{exp} feels the emotion e regarding the character C_{cause}.

Previous relevant studies focus on two tasks, namely social network analysis and sentiment/emotion analysis, both contributing to a computational understanding of narrative structures (see Section 2.4.2). In my thesis, I argue that joining these two tasks may enable new types of social network analysis. I am not aware of any such attempt and therefore propose the task of emotion character network extraction from fictional texts. The task is given a text, a network is to be generated, whose nodes correspond to characters and edges to emotions between characters.

Therefore, in the remaining of my thesis, I will focus on automatic identification of character relationship. Specifically, I will address the following research questions:

Research Questions

Which model is suitable for capturing character relationships?

I explain how I address these questions in the next sections.

5.2 Contributions

My main contributions to this topic are the following:

1. I propose the new task of emotion relationship classification of fictional characters.

2. I provide a fan-fiction short story corpus annotated with characters and their emotion relationships.

I evaluate my models on the textual and the social network graph level and show that a neural model with positional indicators for character roles performs the best. An additional analysis shows that the task of character relationship detection leads to higher performance scores for polarity detection than for more fine-grained emotion classes. Differences between models are minimal when the task is cast as a polarity classification but are striking for emotion classification.

The work done within this topic has potential to support a literary scholar in analyzing differences and commonalities across texts. As an example, one may consider Goethe's *The Sorrows of Young Werther* (Goethe, 1774), a book that gave rise to a plethora of imitations by other writers, who attempted to depict a similar love triangle between main characters found in the original book. The results of this study can potentially

be used to compare the derivative works with the original (see also Barth et al., 2018).

5.3 Corpus

5.3.1 Data Collection and Annotation

An emotion relation is characterized by a triple $(C_{\text{exp}}, e, C_{\text{cause}})$, in which the character C_{exp} feels the emotion e (mentioned in text explicitly or implicitly). The character C_{cause} is part of an event which triggers the emotion e. I consider the eight fundamental emotions defined by Plutchik (2001) (anger, fear, joy, anticipation, trust, surprise, disgust, sadness). Each character corresponds to a token sequence for the relation extraction task and to a normalized entity in the graph depiction.

Using WebAnno (Yimam et al., 2013), I annotate a sample of 19 complete English fan-fiction short stories, retrieved from the Archive of Our Own project[3] (due to availability, the legal possibility to process the texts and a modern language), and a single short story by Joyce (1914) (Counterparts) being an exception from this genre in the corpus. All fan-fiction stories were marked by the respective author as complete, are shorter than 1500 words, and depict at least four different characters. They are tagged with the keywords "emotion" and "relation-

[3] `https://archiveofourown.org`

ships"[4].

The annotators were instructed to mark every character mention with a canonical name and to decide if there is an emotion relationship between the character and another character. If so, they marked the corresponding emotion phrase with the emotion labels (as well as indicating if the emotion is amplified, downtoned or negated). Based on this phrase annotation, they marked two relations: from the emotion phrase to the experiencing character and from the emotion phrase to the causing character (if available, *i.e.*, C_{cause} can be empty). One character may be described as experiencing multiple emotions. The annotated data is freely available[5].

5.3.2 Inter-Annotator Agreement

I calculate the agreement along two dimensions, namely unlabelled vs. labeled and instance vs. graph-level. Table 5.1 reports the pairwise results for three annotators. In the *Inst. labelled* setting, I accept an instance being labeled as true positive if both annotators marked the same characters as experiencer and cause of an emotion and classified their interaction with the same emotion. In the *Inst. unlabelled* case, the emotion label is allowed to be different. On the graph level (*Graph labelled* and *Graph unlabelled*), the evaluation is performed on

[4]Meta-information is enabled for the search functionality.

[5]`http://www.ims.uni-stuttgart.de/data/relationalemotions`

	a1–a2	a1–a3	a2–a3
Inst. labelled	24	19	24
Inst. unlab.	33	27	29
Graph labelled	66	69	66
Graph unlabelled	90	93	92

Table 5.1: F_1 scores in % for agreement between annotators on different levels. a1, a2, and a3 are different annotators.

an aggregated graph of interacting characters, *i.e.*, a relation is accepted by one annotator if the other annotator marked the same interaction somewhere in the text. I use the F_1 score to be able to measure the agreement between two annotators on the span levels. For that, I treat the annotations from one annotator in the pair as correct and the annotations from the other as predicted.

As Table 5.1 shows, agreement on the textual level is the lowest with values between 19 and 33 % (depending on the annotator pair), which also motivated my aggregation strategy mentioned before. The values for graph-labelled agreement are more relevant for the use-case of network generation. The values are higher (66–93 %), showing that annotators agree when it comes to detecting relationships regardless of where exactly in the text they appear.

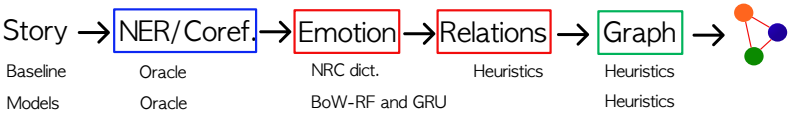


Figure 5.3: Models for the emotion relationship prediction. Oracle: a set of character pairs from the gold data.

5.3.3 Statistics

Table 5.2 on page 138 summarizes the aggregated results of the annotation. The column “All" lists the number of experiencer annotations (with an emotion), the column “Rel." refers to the counts of emotion annotations with both experiencer and cause.

Joy has the highest number of annotated instances and the highest number of relationship instances (413 and 308 respectively). In contrast, *sadness* has the lowest number of annotations with a total count of instances and relations being 97 and 64 respectively. Overall, I obtain 1335 annotated instances, which I use to build and test my models.

5.4 Methods

My model architecture is motivated by the task: using a story as input, I want my model to output a social network of character interactions. One way to implement such a system is to have a pipeline where each component facilitates the final

Emotion	All	Rel.
anger	258	197
anticipation	307	239
disgust	163	122
fear	182	120
joy	413	308
sadness	97	64
surprise	143	129
trust	179	156
total	**1742**	**1335**

Table 5.2: Statistics of emotion and relation annotation. "All" indicates the total number of emotion annotations. "Rel." indicates total number of emotion relationships (including a causing character) instantiated with the given emotion.

Indicator	Implementation example
No-Ind.	Alice is angry with Bob
Role	<e>Alice</e>...<c>Bob</c>
MRole	<e>...<c>
Entity	<et>Alice</et>...<et>Bob</et>
MEntity	<et>...</et>

Table 5.3: Different indicators applied to the same instance. *No-Ind.* means no positional indicators are added. *M* in *MRole* and *MEntity* means that the name of the character is masked. Tag <e> indicates the experiencer. Tag <c> indicates the cause. Tag <et> indicates an entity.

goal. For example, named entity recognizer finds characters, emotion classifier recognizes the emotion, and so forth until the network is generated. The real-world architecture of such a model is shown on Figure 5.3. In this work, I do not use named entity recognition in the first step of the model, as information about character positions is already available in the gold data[6].

I distinguish between *directed* and *undirected* relation prediction. In the directed scenario, I classify which character is the experiencer and which character is the cause, as well as what is the emotion between two characters. For the undi-

[6]I address a real-world application scenario of the model in Chapter 6.

rected scenario, I only classify the emotion relation between two characters. I do not tackle character name recognition here: my models build on top of gold character annotations.

The *baseline* model predicts the emotion for a character pair based on the NRC dictionary (Mohammad and Turney, 2013). It accepts the emotion associated with the words occurring in a window of n tokens around the two characters, with n being a parameter set based on results on a development set for each model.

Further, I cast the relation detection as a machine learning-based classification task, in which each classification instance consists of two character mentions with up to n tokens context to the left and to the right of the character mentions. I compare an extremely randomized tree classifier with bag-of-words features (Geurts et al., 2006) (*BOW-RF*) with a two-layer GRU neural network (Chung et al., 2014) with max and averaged pooling. In the latter, I use different variations of encoding the character positions with indicators (inspired by Zhou et al. (2016), who propose the use of positional indicators for relation detection). My variations are exemplified in Table 5.3 on page 139. Note that the case of predicting directed relations is simplified in the "Role" and "MRole" cases in contrast to "Entity" and "MEntity", as the model has access to gold information about the relation direction.

I obtain word vectors for the embedding layer from GloVe

(pre-trained on Common Crawl, $d = 300$, Pennington et al., 2014) and initialize out-of-vocabulary terms with zeros (including the position indicators).

5.5 Experiments

5.5.1 Experimental Setting

In the classification experiments, I compare the performance of my models on different label sets. Namely, I compare the complete emotion set with 8 classes to a 5 class scenario where I join *anger* and *disgust*, *trust* and *joy*, as well as *anticipation* and *surprise* (based on preliminary experiments and inspection of confusion matrices). The 2-class scenario consists of positive (*anticipation, joy, trust, surprise*) and negative relations (*anger, fear, sadness, disgust*). For each set of classes, I consider a setting where directed relations are predicted with one where the direction is ignored. Therefore, in the *directed* prediction scenario, each emotion constitutes two classes to be predicted for both possible directions (therefore, 16, 10, and 4 labels exist).

The evaluation is performed with precision, recall and F_1 in a cross-story validation setting, in which each story is used as one separate test/validation source. For model selection and meta-parameter optimization, I use 50 % randomly sampled annotations from this respective test/validation instance as a

validation set and the remainder as test data.

Further, I evaluate on three different levels of granularity: Given two character mentions, in the instance-level evaluation, I only accept the prediction to be correct if exactly the same mention has the according emotion annotation. I then aggregate the different true positive, false positive and false negative values across all stories before averaging to an aggregated score (similar to micro-averaging). On the story-level, I also accept a prediction to be a true positive the same way, but first calculate the result $P/R/F_1$ for the whole story before averaging (similar to macro-averaging). On the graph-level, I accept a prediction for a character pair to be correct without considering the exact position.

5.5.2 Results

Table 5.4 on page 144 shows the results on development data and independent test data for the best models. The GRU+MRole model achieves the highest performance with improvement over BOW-RF on the instance and story levels, and shows a clear improvement over the GRU+NoInd. model in the directed 8-class setting. GRU+Role achieves the highest performance on the graph level in the directed 8-class setting. In the undirected prediction setting, all models perform better in the 5-class experiment and 2-class experiment than in 8-class experiment. This is not always the case for the directed predic-

tion, where some models perform better in 8-class experiment (GRU+NoInd., GRU+Entity, BOW-RF).

Figure 5.4 show the performance of models with different number of modeled classes. One may observe that all models perform better in a 2-class scenario (directed and undirected). However, the differences between the models in a 2-class setting are marginal, especially in the undirected scenario. This may suggest that character relations are more nuanced than binary. It also suggests that directionality is an important aspect for the task of relation classification. In the directed classification scenario, the differences between different models are more pronounced, as compared to the undirected scenario.

As expected, we observe a better performance on a graph level for all models, with the highest performance of 47 % F_1 (GRU+MEntity), 63 % F_1 (GRU+MEntity), and 73 % F_1 (GRU+MRole, GRU+MEntity, GRU+NoInd.) in undirected 8-, 5-, and 2-class experiments, respectively, on the development set. In the directed scenario, the highest performances are 41 % F_1 (GRU+Role), 48 % F_1 (GRU+MRole), and 65 % F_1 (GRU+MRole).

Figure 5.5 depicts the performance of all models evaluated on the instance-level for one example run. I tuned the window size parameter on a development set using a set of window sizes of 5, 10, 20, 50, and 100 tokens around character mentions. As one may see, the window size of 5 tokens is the best in the

		Undirected									Directed								
		8 Class			5 Class			2 Class			8 Class			5 Class			2 Class		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Dev Instance level	Baseline	19	31	24	25	38	30	39	100	56									
	BOW-RF	18	18	18	31	31	31	56	56	56	20	20	20	19	19	19	35	35	35
	NoInd.	31	31	31	39	39	39	64	64	64	26	26	26	23	23	23	37	37	37
	Role	19	19	19	35	35	35	55	55	55	33	33	33	34	34	34	57	57	57
	MRole	30	30	30	44	44	44	67	67	67	38	38	38	44	44	44	65	65	65
	Entity	20	20	20	34	34	34	58	58	58	23	23	23	19	19	19	30	30	30
	MEntity	30	30	30	43	43	43	65	65	65	28	28	28	29	29	29	40	40	40
Dev Story level	Baseline	20	32	24	27	39	31	40	100	56									
	BOW-RF	20	24	21	33	36	35	58	59	58	21	25	22	19	23	20	37	39	38
	NoInd.	33	33	33	41	41	41	66	66	66	25	25	25	23	23	23	38	38	38
	Role	19	19	19	34	34	34	55	55	55	33	33	33	35	35	35	56	56	56
	MRole	32	32	32	44	44	44	67	67	67	39	39	39	44	44	44	65	65	65
	Entity	21	21	21	31	31	31	57	57	57	22	22	22	18	18	18	30	30	30
	MEntity	33	33	33	46	46	46	65	65	65	28	28	28	30	30	30	39	39	39
Dev Graph-level	Baseline	36	38	31	50	41	46	88	52	65									
	BOW-RF	68	17	27	72	35	36	70	72	71	72	23	34	79	23	34	54	54	54
	NoInd.	44	44	44	55	55	55	73	73	73	35	35	35	33	33	33	54	54	54
	Role	35	35	35	49	49	49	65	65	65	41	41	41	43	43	43	57	57	57
	MRole	45	45	45	58	58	58	73	73	73	40	40	40	48	48	48	65	65	65
	Entity	37	37	37	50	50	50	68	68	68	39	39	39	29	29	29	49	49	49
	MEntity	47	47	47	63	63	63	73	73	73	39	39	39	39	39	39	52	52	52
Test	MRole Inst.	30	30	30	44	44	44	64	64	64	38	38	38	43	43	43	65	65	65
	MRole Story	33	33	33	45	45	45	65	65	65	39	39	39	43	43	43	66	66	66
	MRole Graph	45	45	45	59	59	59	71	71	71	42	42	42	49	49	49	66	66	66

Table 5.4: Cross-validated results for different models in percentages of precision, recall, and F_1 score. Inst. level: aggregated over all instances in the dataset. Story level: averaged performance on all stories. Graph-level: averaged performance on graph level on all stories. All models with indicators are GRU based.

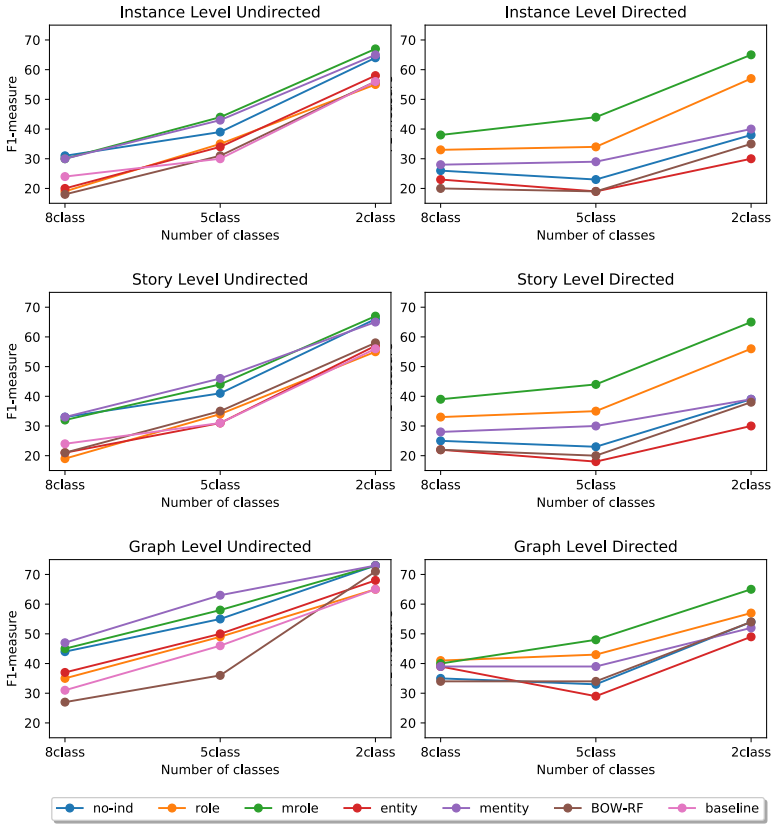


Figure 5.4: Results of experiments with different number of classes.

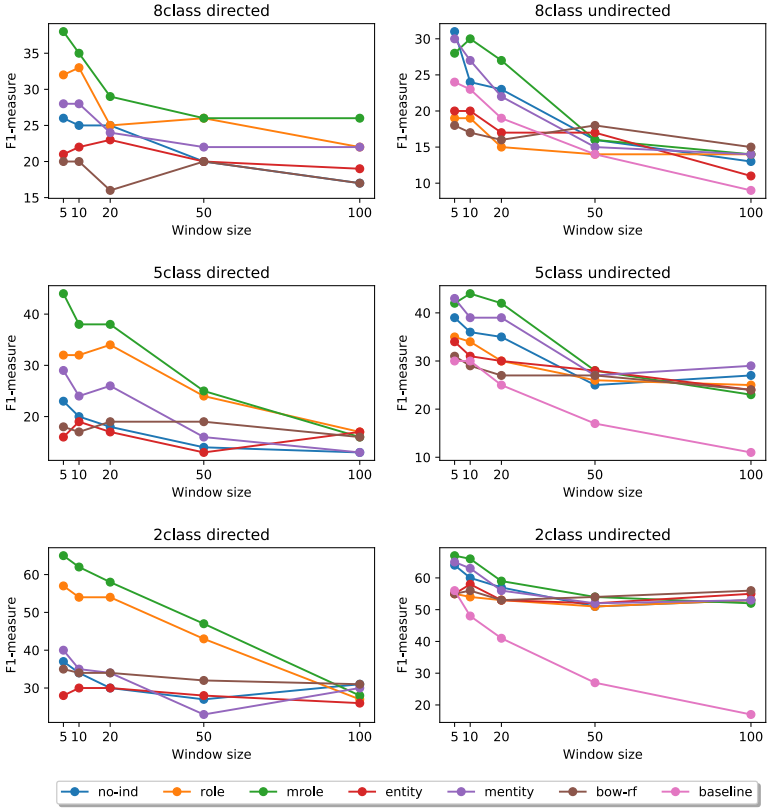


Figure 5.5: Experiment with window sizes.

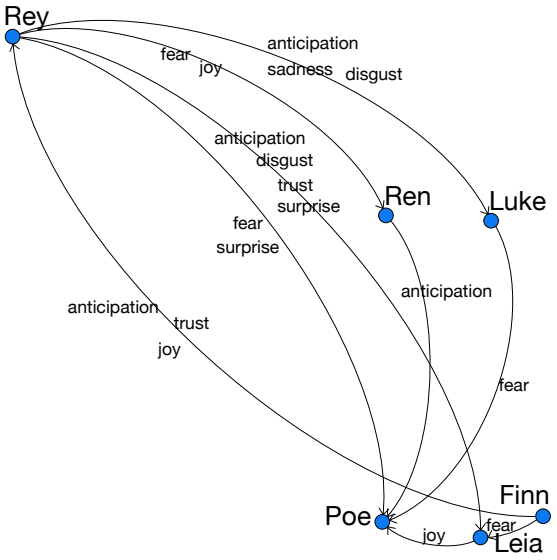


Figure 5.6: Example of a predicted directed network.

majority of cases. The GRU+Entity model shows an exception as it achieves the highest performance with 20 tokens in the 8-class directed scenario. The 2-class GRU+Entity works best with 10 tokens around the character mentions.

Figure 5.6 illustrates a fully predicted network from a fan fiction story based on *Star Wars* universe (Miralana, 2015). The error analysis on the predicted network shows that the mistakes made by our GRU+Role model are not immediately obvious. One example is *trust* relationship between Finn and Rey. Although the textual instance used to classify the inter-

action contains "trust" vocabulary ("they could **help** and be **supportive**"), the overall tone suggests that Finn *anticipates* Rey asking for his help rather than directly imposing trust on her. However, as we do not take into account the exact positions, this mistake is still considered a true positive, as *trust* relationship is present in the gold data. Another example is *anticipation* relationship between Rey and Leia that is tagged with *sadness* in the gold data. Consider the following text that was used to classify the relationship: "She adored the older woman and enjoyed her company . . . , there were certain things that she didn't want to share with her . . . ." The text implies that though Rey is pious towards Leia, some aspects of their relationship do not allow her to be fully open with that woman, hence sadness. The erroneous relationship assignment is then presumably triggered due to specific words such as "adored", enjoyed" and "share", which often indicate *joy* and *anticipation*. This prediction does not count as true positive, as the gold data does not contain *anticipation* among correct relationship between Rey and Leia.

The results show that the sequential and embedding information captured by a GRU as well as additional positional information are all relevant for a substantial performance, at least on the fine-grained emotion prediction task.

5.6 Discussion and Conclusion

In this section, I formulated a new task of emotion character network extraction from fictional texts. Starting from the premise that character emotion interaction can be captured in a graph, I developed a resource annotated with character relationships and proposed several models to automatically classify these relationships. Specifically, in this section I tried to answer the following research questions:

Research Questions

Which model is suitable for capturing character relationships?

My general conclusion to this topic is that *modelling character relationships is a feasible task*. Specifically, I find that a recurrent neural architecture with positional indicators leads to the best results of relation classification (Research Question 1). I also showed that differences between different machine learning models with binary mapping of emotion relation is almost leveled. This may suggest that emotion relation classification is best modeled in a multi-class setting, as emotion interactions of fictional characters are nuanced and do not simply map to either a positive or a negative class (Research Question 2).

However, the results I discussed in this section are based on a setting where names of characters come directly from the annotation. This is an unrealistic scenario, as it is not possible to get character annotations for all books we may be interested in analyzing. Therefore, my models should be evaluated in a real-world application pipeline, in which character pairs are extracted from text automatically using named entity recognition. This calls for an additional research agenda, which I present in the next chapter. Specifically, I will present a pipeline architecture of an end-to-end model that directly predicts the graph from a book. As we shall see, this brings about new research questions that I discuss in detail.

6 Graph Prediction Pipeline

6.1 Introduction

In Section 5, I presented a new task of emotion relationship classification and a neural model that achieves the highest performance on this task. However, all of the models that I presented rely on the annotation of characters. As far as it is very unlikely that character annotations are available for an arbitrary book, here, I turn to a more realistic application-oriented use-case of my model.

Specifically, I develop an end-to-end system that takes as its input a raw text (*e.g.*, a novel), and outputs a social network graph $G = (V, E)$, where nodes $V = (v_1 \dots v_n)$ are characters appearing in the story and edges $E = (e_1 \dots e_n)$ are emotions that hold between any two nodes.

There are several considerations related to this goal. First of all, for a better evaluation of the system it is desirable to include some variability in terms of the type of texts the system analyzes. My original models were trained and evaluated on a corpus of fan fiction. However, in order to scale the mod-

els to any unseen text, they should work on different genres. One approach to solve this problem involves the use of additional training data, which comes from a more general type of fiction (*e.g.*, classical fictional books) than fan fiction. A complementary solution is to include an additional test set, specifically books from different genres, and evaluate the new model on this test set. Each of these techniques is associated with certain issues which I discuss in detail below.

The research question I will address in this section is, therefore:

Research Questions

How to manipulate a precision-recall tradeoff of the social network generation pipeline that uses NER and emotion recognition components?

I will present the pipeline in the succeeding sections.

6.2 Contributions

My contributions related to this topic are the following:

1. I present a pipeline that takes a book as its input and outputs a graph that reveals the predicted emotions be-

tween characters.

2. I show that different strategies for edge prediction lead to different results at the output of the pipeline.

3. I provide an analysis of the pipeline performance, which helps in a better understanding of improvements to allow for higher precision and coverage of the pipeline.

4. I evaluate my system on out-of-domain data.

6.3 Data Collection and Annotation

I opt for a trade-off between genre variability and annotation granularity in the pipeline. I tackle the problem of unfeasible overhead required to annotate dozens of books on a low level (*e.g.*, annotating all relevant characters in text with emotions) by turning to book plot summaries and using them to create gold data (high-level annotation of character relationship). The advantage of using plot summaries as a basis for relationship annotation has been used before, *e.g.* by Chaturvedi et al. (2016) and Srivastava et al. (2016), who use books summaries for inferring evolving inter-personal relationships of literary characters. I will use the gold data only to evaluate the performance of the pipeline and not to train the system.

This approach has both advantages and disadvantages. An advantage of using the summaries is that they are usually short

and include relevant information about the plot, which results in less annotation effort. However, this comes with a disadvantage of being detached from text, in terms of the exact character offsets, which makes direct evaluation on the text level not possible. I explain how I solve this problem in the upcoming sections.

6.3.1 Corpus

As a source of book summaries I use study guides for students published by SparkNotes[1]. SparkNotes is a resource that helps school students better understand books. For that purpose, the website contains specially written study guides for required school literature and other popular books. Each study guide contains a plot overview, a character list with a profile description of each character, and a discussion of main themes and ideas expressed in a book.

Table 6.1 on page 179 summarizes the corpus of plot overviews I collect. Overall, I collect 26 book summaries from SparkNotes along with the genre information for each book.

Each plot summary is then annotated by me using a plain text editor. The annotation was done as follows: I read a summary and while reading, I decide whether two or more characters mentioned in the summary experience emotions towards each other. If yes, I record the information about their

[1] `https://www.sparknotes.com/lit/`

relationship.

I use Plutchik's set of emotion labels, namely *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*. Additionally, I include *other-emotion* label, similarly to the annotation presented in Chapter 4, to account for complex emotion relations, such as *shame, jealousy, guilt* and *pride* and others. Whenever possible, I map such "traditional" interpersonal relations as *love*, *friendship*, *rivalry*, etc. to Plutchik's set of emotions (*e.g.*, *love* to *joy*). The choice in such cases was mostly intuitive and relied on a greater context given in a summary.

Table 6.2 on page 180 exemplifies the final annotation. Every pair of characters may be involved in more than one emotion relationship, as these relationships evolve throughout the story and this fact may or may not be reflected in a summary.

In parallel, I prepare a corpus of real books, for which the relationships were annotated with a help of book summaries. I will use these texts to *test* the pipeline. To that end, I scan hard copies of the titles in Table 6.1 (page 179) and convert them to plain text using OCR software, Abbyy FineReader Pro[2], v.12.0.1. As one can see from the table, I select books from different genres and time periods. This enables pipeline evaluation on as diverse data as possible. I describe the pipeline in the next sections.

[2] `https://www.abbyy.com/en-eu/finereader/`

6.3.2 Annotation Statistics

The results of the annotation are shown in Table 6.3 on page 180. There are 259 characters that are involved in emotion relationships overall. Together they are involved in 735 emotion relationship instances, meaning that average ratio of emotion relations per pair is 2.8.

Joy dominates by the number of annotations and is followed by *anger* and *trust*. Conversely, *other emotion, fear* and *disgust* represent have the least number of annotations.

These observations may suggest that *joy*, *anger* and *trust* are the most representative in fiction and play dominant role in the evolution of character relationships and plot development.

I plot a heatmap showing the normalized frequency of emotion co-occurrence with genre (Figure 6.1). This heatmap can be interpreted as emotion-genre association. Each cell is normalized by overall number of emotion interactions annotated within a given genre.

As the heatmap shows, *joy*, *anger*, and *trust* relations again are on top of the list of the most frequent emotions. *Joy* is most strongly associated with the *coming-of-age* fiction, *gothic fiction*, and *novel of manners*, *Anger* has the strongest associations with *satire* and *allegory*. *Trust* is more common for *psychological fiction*.

Some cells show close to zero emotion-genre association values. In this analysis, it cannot be interpreted as a lack of as-

sociation, because these genres are singletons (*autobiography, gothic fiction, and love story*, with just one instances in the dataset each.

The analysis of the heatmap along a *genre* axis, shows that *coming-of-age* books are associated with *joy* and *anger* relationships; *fantasy* depicts *joy, anger* and *trust* interactions; and *science fiction* portrays *joy, anger* and *sadness* and *surprise*.

6.4 Pipeline

I now turn to describing the architecture of my pipeline that predicts an emotion-informed graph from an input book. Figure 6.2 depicts the architecture of the pipeline. The pipeline consists of the following components.

First, the NER/COREF component based on spaCy[3] (Honnibal, 2013) and Stanford NER (Finkel et al., 2005) finds all mentions of characters in the text together with their coreference mentions[4]. This information is used in the next step (RELATION PROCESSOR), which generates the pairs of interacting characters and collects the contextual information about their interaction[5]. In the next steps, the classification com-

[3] `https://spacy.io`

[4] `https://github.com/huggingface/neuralcoref`

[5] The characters are considered to be in interaction based on the `distance parameter` $\text{dist}_X(p, q)$, where p and q are the strings corre-

ponent (CLASSIFIER) classifies textual instances that characterize the relationship between two characters and outputs an emotion-labelled graph (GRAPH BUILDER), where nodes are characters and edges connecting them are labelled with emotion relations provided by the classifier from the previous step.

6.5 Evaluation

To understand the effect each component has on the final result, I evaluate the performance of the pipeline along four different dimensions: 1) NER, 2) relation preprocessing, 3) emotion classification, 4) complete graph with nodes and edge labels. I describe each evaluation in detail below.

6.5.1 NER

The goal of this evaluation step is to understand what is the potential for finding the relevant pairs of characters. It can be helpful in understanding of error propagation in the pipeline.

sponding to these characters and X is the number of tokens between them. In addition, I use the `context parameter` $\text{cont}_Y(p, q)$, where p and q are the strings corresponding to these characters and Y is the number of tokens before the character p and after the character q. While the former measure allows for detecting those characters that are closely related to each other, the latter one enables a contextual analysis of their relationship.

I evaluate the NER performance by comparing two sets of character names: one is a gold set based on the annotation, and one is a predicted set based on the NER results. If the predicted name of a character appears in the gold data, regardless of the textual position, I consider it a true positive.

6.5.2 Relation Processor

Evaluation of relation preprocessing allows to understand the potential of finding the relevant pairs of characters, and therefore, is directly linked to the next step in the pipeline, which is called `CLASSIFIER`. A predicted relation is considered a true positive if it appears in the gold data. For example, if the gold data contains "Alice–Bob" pair and the same character pair also appear in the predictions, then their interaction is considered a true positive. Note that this evaluation step does not consider an emotion attached to the relationship. It considers a mere fact that two characters are in relationship.

6.5.3 Classifier

The goal of this evaluation step is to estimate an upper-bound for overall correct emotion label assignment. Specifically, it shows how well does emotion classification model work. Again, knowing the upper-bound can help in the error analysis of the final results. Note that this evaluation step does not take into

account character names associated with an emotion-relation, but operates on emotion-level only.

6.5.4 Final Aggregated Evaluation

Finally, I evaluate the pipeline on an aggregation of all its steps, from NER to emotion-relation classification. This evaluation step assesses the performance of the pipeline on the final output, an emotion labelled social network of characters. Thus, this is a realistic evaluation of the pipeline performance. Specifically, each prediction–a pair of characters with attached emotion to their relationship–is tested for correctness against the gold data. If the predicted pair appears in the gold data and has the same emotion label as a predicted one, then the prediction is considered a true positive.

6.6 Model Training

For the `CLASSIFIER` component in the pipeline I use the model presented in Section 5.4. Namely, a two-layer GRU model with a combination of average and max pooling between recurrent layers. I use two data sources to re-train the model.

Source 1: Fan fiction. The fan fiction stories presented in Section 5.3. However, the stories are not split into train and test sets, but the entire corpus is used to train the model.

Source 2: Additionally annotated sentence triples.

This data comes from the spin-off of REMAN annotation project described in Chapter 4. The working name of the project is REMAN`en`, where `-en` stands for *entity* and reflects the entity-oriented annotation. The only difference of REMAN`en` from original REMAN is that only entities (characters) can take roles of emotion *cause* and *target*, which is similar to the annotation logic used in a fan fiction dataset. The source of the texts is Project Gutenberg.

Both sources are merged into a single dataset that is used to train the model.

Table 6.4 on page 182 summarizes the properties and quantity of the training data. Column *Fan fic* summarizes the fan fiction corpus, column *REMAN`en`* summarizes the additional data sampled from Project Gutenberg and annotated with character relationships.

Overall, there are 2590 training instances. Each instance is a pair of characters with emotion between them. There are nine emotion labels in total (eight Plutchik's emotion and additional "other emotion"). *Joy* has the largest number of instances (566 in total), while *other emotion* is the least represented class with 119 instances.

The model is trained with No-Ind. representation (see Table 5.3 on page 139) for 20 epochs and the model is saved to be used in the pipeline.

6.7 Experiments and Setup

In this section I describe the experimental setup that I use with the pipeline.

6.7.1 Estimation of Raw Potential

With this experiment I aim to estimate the raw potential of both each individual component and the overall pipeline performance. To that end, I perform a single pipeline run with default parameters on all 26 titles in the dataset, evaluating the performance of each component, as well as the aggregation of all components (see Section 6.5). Specifically, I evaluate `NER/COREF`, `RELATION PREPROCESSOR`, `CLASSIFIER`, and `AGGREGATION` components. The goal of the experiment is to get an overall understanding of interactions between different components that affect the final output, a predicted graph.

6.7.2 Hyperparameter Optimization via Cross-Validation

Each component has a number of associated parameters and these parameters may influence the subsequent components. Therefore, I optimize the parameter values. To determine the optimal values for each component in the pipeline, I perform a grid search over parameter space in a cross-validated fashion.

In particular, I first split the dataset into ten folds. I use

nine folds to optimize the parameter values and test the model with the optimal parameters on the remaining fold.

Specifically, I optimize the following parameters:

RELATION

- `Minimal number of interactions of NEs`. If two NEs of type `PESON` interact[6] less than a threshold value, the pair is excluded from the relationship graph. Search space ranges from 1 to 5, 10, 15, and 20.

- `Distance parameter`. Distance in tokens between two NEs of type `PERSON`. If distance $\leq$ threshold value, a pair of characters is considered to be in an interaction. The text between two NEs is used for the relationship classification. Search space includes values of 5 to 10, 15, and 20.

- `Context parameter`. Number of tokens to the left and to the right of NEs. The text to the left and to the right is used for the relationship classification. Search space includes values of 5 to 10, 15, and 20.

- `Emotion filter`. When applied returns only those relationship instances that contain emotion-associated terms. Search space ranges over boolean values.

[6]The interaction is determined by a `distance parameter`.

CLASSIFIER

- `Confidence`. Confidence of prediction. If confidence of emotion relationship prediction is less than a threshold value, then prediction is discarded. I consider threshold values from 0.1 to 0.9.

6.7.3 Results and Discussion: Estimation of Raw Potential

First, I estimate the raw potential of my pipeline. Table 6.5 on page 183 shows the results of this experiment. I present and discuss the results below.

Results

Column *NER/COREF* reports the performance metrics for the `NER/COREF` component. It is an estimate of finding literary characters that are relevant for emotion relationship classification. The lowest precision and recall of 18% and 30% respectively are observed for *The Tale of Two Cities*. The highest precision of 97% is achieved on *Divergent*, while the highest recall of 99% is observed for *Little Women*. It can be seen that on average recall is higher than precision by 8 pp and the average F_1 score is 74%.

Column *RELATION* shows the performance of the `RELATION` component. It shows an estimate of how well can the pipeline

put characters in a relationship using a `distance parameter`. The first striking observation is that performance for some of the books is zero (*A Clockwork Orange*, *The Hunger Games*). The `RELATION` component shows mixed predictions, with a tendency to higher recall than precision. The highest precision of 67% is achieved on *The Lion, The Witch, and The Wardrobe*. Recall is on average higher than precision by 29 pp and the average F_1 score is 38%.

The *CLASSIFIER* column shows the performance of the `CLASSIFIER` component and estimates the raw potential of the classifier to recognize emotions without associating them to interactions of characters. As one can see, the `CLASSIFIER` component achieves average recall of 99% and average precision of 77%.

Finally, the *AGGREGATION* column shows the result for the final step in the pipeline, where the system is evaluated in an aggregation of all components. It can be seen that the results are mixed with lowest performance metrics being zero and the highest precision and recall of 19% and 56% respectively (*Never Let Me Go*).

On importance of recall in evalutation

Consider a case in which NER fails to find an entity that is a part of an emotion relation (given in the gold data). This means that the whole emotion relation in gold data cannot be found and classified by the pipeline, as one or more of its parts were not found in a preceding step.
On the other hand, entities that are detected by NER but not in the gold data, are not necessarily false positives. As gold data is obtained from book summaries, it can be the case that some mentions of characters were omitted by summary authors, as irrelevant or not important for the plot. However, one may not argue that an unimportant pair of characters is not involved in an emotion relationship just because the summary does not mention it.
Therefore, in the evaluation, I propose to *regard recall higher than precision*, as 1) we are interested in finding as many relevant emotion-relation pairs as possible, 2) emotion-relation pairs that are not mentioned in the gold data, are not necessarily incorrect, but could have been omitted due to their lesser importance for the plot in the eyes of a summary author.

Discussion

The experiment shows that the task of correctly predicting character relationships on a full-text level is a difficult one, with the average precision, recall, and F_1 score of 7%, 21%, and 9% respectively. When evaluated independently, different components, namely `NER/COREF` , `RELATION` , and `CLASSIFIER` , show decent performance, especially in terms of completeness of predictions. Strikingly low performance of a complete pipeline, evaluated at the aggregation level, is presumably related to error accumulation and propagation.

Consider the pipeline prediction for *Flowers for Algernon*. As the results show the `NER/COREF` component finds 90% of relevant character names and is 73% exact in its predictions. Put simply, that means that `NER/COREF` is able to find most of the character names that are in the gold data (high recall), but also predicts entities that are not in the gold data (lower precision).

Next, the set of predicted character names is passed down the pipeline, to the `RELATION` component. The `RELATION` component does not predict new named entities. It uses a `distance parameter` to put the entities it received from the `NER/COREF` component into a relationship. Assuming that F_1 score for NER for *Flowers for Algernon* is 81%, we cannot expect that the `RELATION` component can overreach this upper-bound, as the number of correctly predicted pairs of characters will de-

pend upon whether all relevant characters were detected during NER step.

Then the `RELATION` component predicts the character pairs that are interacting and in doing so makes mistakes. The precision, recall, and F_1 score now are 20%, 56%, and 30% respectively (46 pp below `NER/COREF` on average). What exactly these numbers suggest? Remember that the `NER/COREF` component 1) did not find all the relevant character names, and 2) predicted entities that are not in the gold data. This means that:

- a great deal of the relation pairs were false positives because either one or two characters forming the pair were not in the gold data (hence low precision);
- some of the relation pairs that should have been found were not found due to 1) not optimal heuristics for relation assignment, or 2) due to the fact that relevant characters were not predicted by the `NER/COREF` component (hence low recall).

At this step, we know that even though the `CLASSIFIER` is able to make correct predictions over textual instances in principle, the error propagation leads to low overall pipeline performance, leading to precision, recall, and F_1 score of 5%, 13%, and 7% respectively, which is 73 pp below the potential theoretical upper-bound. The predicted network, together

with the gold network, is depicted on Figure 6.3. It can be seen that the predicted network is abundant with false positives (according to gold data, but not always according to actual book; see discussion "On importance of recall in evaluation" on page 166).

A more extreme example of error propagation is *Hunger Games*. With precision, recall, and F_1 score of 37%, 67%, and 48% in the `NER/COREF` component, the number of correctly predicted relation pairs is zero, which results in final F_1 of 0%.

These limitations may presumably be related to non-optimal parameters used with each component. Because it is unlikely that the same parameter set is going to be optimal for an arbitrary book, I optimize and test the parameters in cross-validation fashion. The results of this experiment are reported in the next section.

6.7.4 Results and Discussion: Cross-Validation

Table 6.6 on page 184 reports the performance of the pipeline obtained with a 10-fold cross-validation. As a reminder, within each run, the optimal parameter set was determined on nine folds, and the most optimal parameters were then applied to the remaining test fold. Figure 6.4 (page 185) summarizes the search space for each parameter. Due to ties[7] in the returned

[7] A tie occurs when multiple various parameter sets lead to identical performance on a training set.

parameters, the model was applied to each book within a fold multiple times with different optimal parameters. Column *runs* shows the number of times the model was applied to a single book within a test fold.

It can be seen that the system leans towards higher recall. The average recall across ten folds is 30% and is higher than precision by 17 pp. The average precision is 13% across ten folds on average.

For some books, the performance metrics are zero, which partially results in rather spread-out standard deviation values of 11.68, 26.45, and 13.86 pp for precision, recall, and F_1 correspondingly.

Column *stddev* shows standard deviation values for F_1 score within a fold. It can be seen that, for some books, multiple different parameter sets do not result of skewed performance, which is evidenced by rather low standard deviation (*e.g.*, standard deviation of 1.63 across four runs for *The Lion, The Witch, and The Wardrobe* with F_1 of 29%). This may suggest that these books are relatively easy for the model and the final performance is not conditioned on a parameter set. On the other hand, some books, such as *A Storm of Swords*, show both low F_1 score and low standard deviation values across multiple runs, which may be an indicator that such books are especially difficult for the model and are not sensitive to parameter optimization.

Performance and Network Properties

Given the considerations regarding the model outlined above, I carry out an analysis of the networks predicted by the model for each model. Specifically, I calculate the following statistical network metrics:

Density The ratio of actual edges in the network to all possible edges. Values for this metric range between 0 and 1, where 0 indicates that the network has no edges, and 1 that the network is fully connected.

Diameter Evaluates the size of the network. Calculated as the size of the path between two nodes that are the furthest apart.

Triadic closure A structural measure. Supposes that if two characters interact with one of the characters, they are likely to interact with each other. Outputs a number between 0 and 1, where 0 indicates that there are no triangles in the network, and 1 that the network has the maximum possible number of triangles.

Degree A sum of node edges. Measures node importance. The higher the degree of a character, the more characters

he/she interacts with, and, hence, the more important the character is.

Betweenness centrality Measures how often a node serves as a mediator on a shortest path for two disparate nodes. Measures node importance based on the fact that the nodes serve as bridges between groups of nodes. Outputs a number between 0 and 1.

I estimate the influence of network properties on performance by calculating Spearman correlation between them and F_1 score. The resulting properties of predicted networks and corresponding correlation values are summarized in Tables 6.7 on page 186 and 6.8 on page 187. Due to ties and multiple runs of the model on each book, the reported statistical metrics are averaged across runs.

It can be seen from the Table 6.7 that density is in orthogonal relationship with performance (Spearman's rho = 0.00) meaning that this network property has no influence on the performance. At the same time, diameter, triadic closure, degree, and betweenness centrality show low positive correlation (between 0.21 and 0.40) suggesting that the higher the complexity of the networks are the lower the performance in terms of F_1 score. For example, Spearman's rho of 0.21 for diameter, that is the size of the network, can be interpreted as indicating that performance for large networks is low because there are

more possibilities for the model to make a mistake. This is partly supported by the inverse correlation of −20 between book length and F_1 score (not shown in Table 6.7), which means that the longer the book is, the more difficult it becomes to find the relevant relationships and classify them correctly. The same applies to all other network properties such as triadic closure, degree, and betweenness, which essentially measure the complexity of the network. Therefore, the Spearman's rho values may support the idea that the more complex the network is, the lower the performance of the model.

Some support for this argument can be found in Table 6.8 (page 187 that shows the correlation of network properties with each other. It can be seen that some correlating network properties display close correlation values to F_1. For example, diameter with average density and triadic closure with average density show correlation of 0.62 rho. Likewise, all of the three metrics *positively* correlate with F_1 on a scale from 0.21 to 0.35 and a standard deviation of 7.3.

Performance and Parameter Sets

Figure 6.5 on page 187 shows how *final aggregated performance*[8] change with different parameters. Each plot shows the effect that the change in the parameter has on performance.

It can be seen that for the *interaction* parameter, that is the

[8] See section Final Aggregated Evaluation.

threshold number of interactions between two characters, the recall is highest at 1, precision is highest at 20, and the optimal F_1 score is achieved at 15 and 20 interactions, meaning that the more often two characters interact, the more likely it is that the model will make a correct prediction. A presumable reason is that character pairs with interaction rate at higher threshold values are more likely to be in the gold data than pairs with low interaction rate.

For the *distance* parameter, that is a minimal number of tokens between two characters to consider them as an interacting pair, the recall is highest at 20, precision is highest at 6, and F_1 score is highest at 10, 15, and 20. For the *context* parameter, which tells the classifier how many context words to the left and to the right of the characters should be used, the highest F_1 score is obtained at values 7 and 20.

Finally, the *classifier* parameter specifies the classifier confidence in the emotion prediction. One can see that precision is highest at a confidence of 0.9, recall and F_1 are the highest at 0.3[9]. This supports out intuition that the more restrictive the acceptance of a prediction is, the more likely it is that the prediction is of high quality. However, this also means that the number of such high quality predictions is extremely low leading to very low recall and, hence, F_1 score.

[9]Again, note that performance is measured on aggregated level, not only for emotion prediction.

Performance Evaluated with Binary Mapping

As an additional analysis, I estimate the model potential by mapping predicted emotions to binary values, namely *anger, fear, disgust, sadness* and *other emotion* to a negative class, and *trust, joy, anticipation*, and *surprise* to a positive class.

It can be seen from Table 6.9 on page 189 that binary mapping favors the final performance, with average precision, recall, and F_1 score being almost twice as higher compared to non-binary mapping. The standard deviation between different folds also grows noticeably, which is explained by the fact that though for many books the performance almost doubles, for other books it stays at zero values.

Overall, the analysis shows that polarity mapping could be a feasible option that can replace fine-grained analysis in situations where an overall impression about character relations suffice. For example, Figure 6.6 on page 188 shows an example of network for *Northanger Abbey* with binary relations. It can be seen that there are more negative than positive relations in the book. However, the network is missing some important nodes and edges, hence F_1 score of 63.

Discussion

Overvall, I observe that the results obtained on the task of emotion relation classification are rather mixed. It can be seen

from Table 6.6 (page 184) that the highest F_1 scores achieved by the model is 43%, 42%, and 33% (*Never Let Me Go*, *Lord of the Flies*, and *Northanger Abbey* respectively). The results for *Never Let Me Go* and *Lord of the Flies* show substantial recall of 71% and 93% respectively, but rather low precision of 31% and 27%. I argue again that recall should carry more weight than precision in my evaluation, as it is more important to find as many character pairs as possible. However, to enable the use of my model in qualitative literary analysis, it is important to have substantial precision as well, which is not the case due to large number of false positives. Presumably, this happens due to pair assignment heuristics that searches for character occurrences and pairs them using distance parameter.

There are seven titles in the dataset for which the model cannot beat the zero F_1 score. These are *The Hunger Games*, *Flowers for Algernon*, *Divergent*, *Absalom, Absalom!*, *A Tale of Two Cities*, *A Farewell to Arms*, and *A Clockwork Orange*. A presumable reason for that is that these books portray rather complex character interactions and have a great number of important characters which is not captured by my pair assignment heuristics.

At the same time, binary mapping of emotions leads to a less penalizing evaluation, which explains higher performance numbers. Thus, F_1 score for *Never Let Me Go*, *Lord of the Flies*, and *Northanger Abbey* grows to 75%, 67%, and 63%

respectively.

In general, I believe that the task still remains feasible, especially with binary mapping of emotions. I see the main bottleneck in high number of false positives, which result in low precision.

6.8 Genre Classification

In the remaining part of the section I describe an experiment on genre classification on the set of books I used for character relationship classification.

6.8.1 Motivation

With this experiment I aim to test if the *predicted emotion-informed networks* can be used as features for a genre classification task. The motivation behind this analysis is to understand whether the accrued groundwork I have done within this thesis could be used for a downstream task, such as genre analysis. My curiosity is not idle: If the networks predicted by my model can indeed be used as features, it means that the model captures important information such as genre of a book. This, in turn, would motivate further research and, perhaps, shed more light on the results of emotion relation classification. Specifically, it can help to estimate whether network properties are indeed an important aspect. Note that the genre experi-

ment described in this section is not related to the experiments described in Section 3.

I describe the methods and features I use in detail in the next section.

6.8.2 Features and Experimental Setup

To estimate if the predicted networks encode and retain genre information I define three sets of features.

Network features Statistical features capturing structural network properties. Namely, density, diameter, triadic closure, average degree, and average betweenness.

Emo n features Most common emotions associated with n most important characters. The most important characters are defined by a degree (that is number of connections to other characters).

Network+Emo n features Concatenation of Network and Emo n features.

Title & Year of Publication	Author	Genre	Coarse label
Absalom, Absalom! (1936)	W. Faulkner	family drama	general fiction
A Clash of Kings (1998)	G.R.R. Martin	fantasy	fantasy
A Clockwork Orange (1962)	A. Burgess	satire	general fiction
A Farewell to Arms (1929)	E. Hemingway	realism	general fiction
A Game of Thrones (1996)	G.R.R. Martin	fantasy	fantasy
A Storm of Swords (2000)	G.R.R. Martin	fantasy	fantasy
A Tale of Two Cities (1859)	C. Dickens	historical fiction	general fiction
Anna Karenina (1877)	L. Tolstoy	realism	general fiction
Cat's Cradle (1963)	K. Vonnegut	sci-fi	sci-fi
Dandelion Wine (1957)	R. Bradbury	autobiography	sci-fi
Divergent (2011)	V. Roth	sci-fi	sci-fi
Emma (1815)	J. Austen	novel of manners	general fiction
Fahrenheit 451 (1953)	R. Bradbury	sci-fi	sci-fi
Flowers for Algernon (1959)	D. Keyes	sci-fi	sci-fi
Jane Eyre (1847)	C. Brontë	gothic fiction	general fiction
Little Women (1868)	L. M. Alcott	coming-of-age fiction	general fiction
Lord Jim (1900)	J. Conrad	modernism	general fiction
Lord of the Flies (1954)	W. Golding	allegory	general fiction
Mansfield Park (1814)	J. Austin	family drama	general fiction
Never Let Me Go (2005)	K. Ishiguro	sci-fi	sci-fi
Northanger Abbey (1817)	J. Austen	novel of manners	general fiction
Pride and Prejudice (1813)	J. Austen	novel of manners	general fiction
The Great Gatsby (1925)	F. S. Fitzgerald	realism	general fiction
The Hunger Games (2008)	S. Collins	sci-fi	sci-fi
The Lion, The Witch, and The Wardrobe (1950)	L. Caroll	fantasy	fantasy
The Unbearable Lightness of Being (1984)	M. Kundera	love story	general fiction

Table 6.1: Books for which summaries were collected.

character1	character2	emotion
Stannis	Joff	anger
Renly	Joff	anger
Robb	Theon	trust
...	...	...

Table 6.2: Relationship annotation example.

Emotion relations	Count
anger	163
anticipation	62
disgust	32
fear	31
joy	208
other-emotion	35
sadness	73
surprise	49
trust	82
overall relations	735
overall characters	259
average relations per pair	$\backsimeq$ 2.8

Table 6.3: Emotion relationship annotation summary.

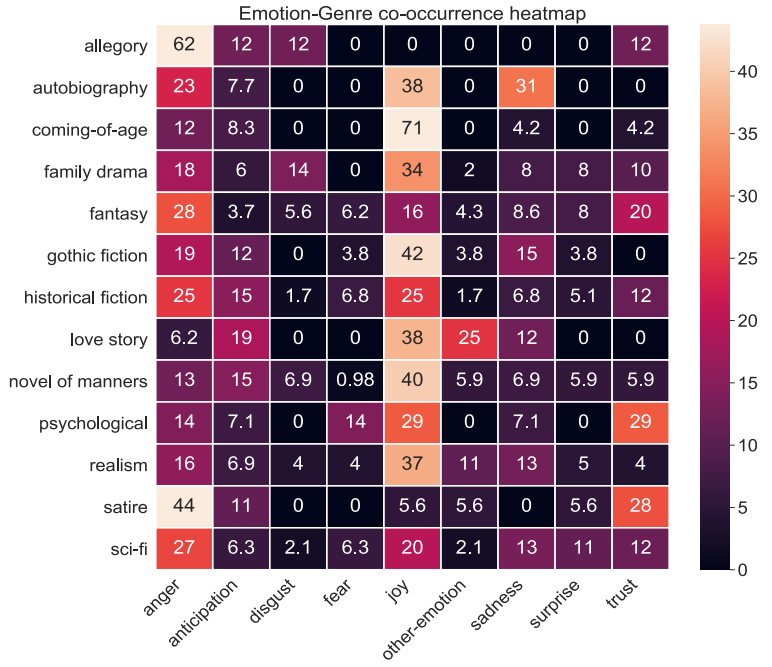


Figure 6.1: Heatmap showing frequency of emotion relationships being annotated with certain genres. Each cell shows the percentage each emotion has in a respective genre (normalized by the overall number of emotion relations annotated for a given genre).

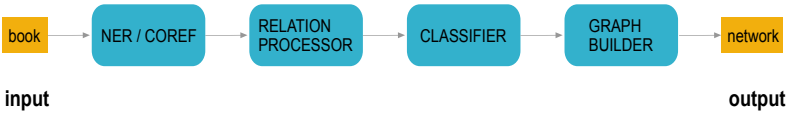


Figure 6.2: Pipeline architecture.

Emotion relations	Fan fic	REMANen	Total
anger	197	132	329
anticipation	239	162	401
disgust	122	140	262
fear	120	107	114
joy	308	258	566
other-emotion	0	119	119
sadness	64	140	204
surprise	129	50	179
trust	156	147	303
overall relations	1335	1255	**2590**

Table 6.4: Data used for re-training the fan-fiction emotion-relation model presented in Chapter 5.

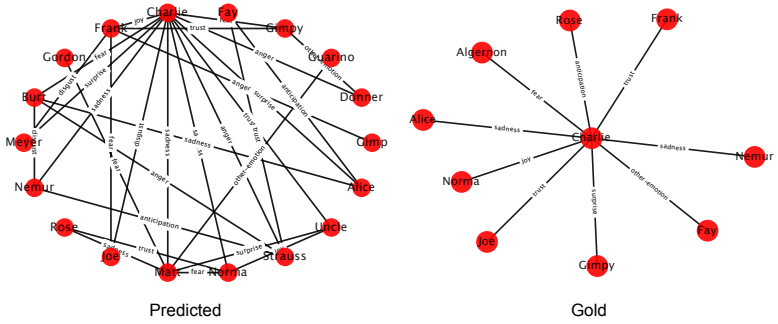


Figure 6.3: Predicted network for *Flowers of Algernon* vs. gold.

	NER/COREF			RELATION			CLASSIFIER			AGGREGATION		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
A Clash of Kings	42	96	57	4	47	8	77	100	87	1	12	1
A Clockwork Orange	74	44	56	0	0	0	64	90	75	0	0	0
A Farewell to Arms	95	57	72	20	13	16	56	83	67	0	0	0
A Game of Thrones	82	96	89	14	60	23	88	100	94	1	10	2
A Storm of Swords	86	86	86	22	53	31	88	100	94	3	11	4
A Tale of Two Cities	18	30	23	5	10	7	78	100	88	0	0	0
Absalom, Absalom!	39	94	55	8	44	14	65	100	78	1	6	2
Anna Karenina	84	89	86	47	83	60	100	99	100	12	43	19
Cat's Cradle	77	79	78	38	56	45	100	100	100	5	10	7
Dandelion Wine	54	87	67	42	77	55	45	100	62	6	27	10
Divergent	97	80	87	22	30	26	80	100	89	0	0	0
Emma	77	99	87	32	89	47	88	100	94	8	27	12
Fahrenheit 451	86	68	76	56	48	51	67	100	80	6	5	5
Flowers for Algernon	73	90	81	20	56	30	90	100	95	5	13	7
Jane Eyre	73	98	84	19	100	32	75	100	86	3	17	5
Little Women	77	99	87	26	97	41	56	100	72	3	48	5
Lord Jim	59	75	66	31	47	37	69	100	82	0	0	0
Lord of the Flies	80	88	84	35	89	50	46	100	63	15	78	25
Mansfield Park	77	93	84	27	86	41	91	100	95	7	27	12
Never Let Me Go	62	89	73	55	89	68	56	100	72	19	56	28
Northanger Abbey	72	93	81	34	76	47	89	100	94	11	15	12
Pride and Prejudice	96	89	93	49	76	60	100	100	100	16	25	19
The Great Gatsby	94	78	85	62	70	65	100	100	100	12	14	13
The Hunger Games	37	67	48	0	0	0	79	100	88	0	0	0
The Lion, The Witch, and The Wardrobe	91	82	87	72	82	77	78	100	88	19	34	24
The Unbearable Lightness of Being	78	38	51	67	35	46	56	100	71	11	6	8
AVERAGE	**73**	**81**	**74**	**31**	**58**	**38**	**76**	**99**	**85**	**6**	**19**	**8**

Table 6.5: Evaluation of separate pipeline components. No grid search is applied.

	P	R	F_1	runs	stddev
A Clash of Kings	2	13	3	5	2.05
A Clockwork Orange	0	0	0	6	0
A Farewell to Arms	0	0	0	4	0
A Game of Thrones	3	25	6	6	2.88
A Storm of Swords	3	16	5	5	1.02
A Tale of Two Cities	0	0	0	2	0
Absalom, Absalom!	0	0	0	6	0
Anna Karenina	15	62	24	6	1.7
Cat's Cradle	20	23	21	5	5.74
Dandelion Wine	13	45	20	4	2.17
Divergent	0	0	0	6	0.37
Emma	22	48	30	4	2.16
Fahrenheit 451	24	22	22	5	7.99
Flowers for Algernon	0	0	0	4	0
Jane Eyre	10	38	16	4	3.26
Little Women	5	76	10	2	0
Lord Jim	14	31	19	5	5.85
Lord of the Flies	27	93	42	2	3
Mansfield Park	13	40	19	5	3.97
Never Let Me Go	31	71	43	6	5.27
Northanger Abbey	30	39	33	6	4.84
Pride and Prejudice	20	50	28	5	4.58
The Great Gatsby	38	28	31	6	14.38
The Hunger Games	0	0	0	5	0
The Lion, The Witch ...	21	48	29	4	1.63
The Unbearable Lightness ...	19	14	16	5	11.13
AVERAGE	13	30	16	4.7	
stddev	11.68	26.45	13.86		

Table 6.6: The final pipeline performance achieved via 10-fold cross-validation. Each book within each test fold was classified n times, with n ranging from 2 to 6 (due to ties in metaparameter optimization), hence *steddev* for each individual book is given.

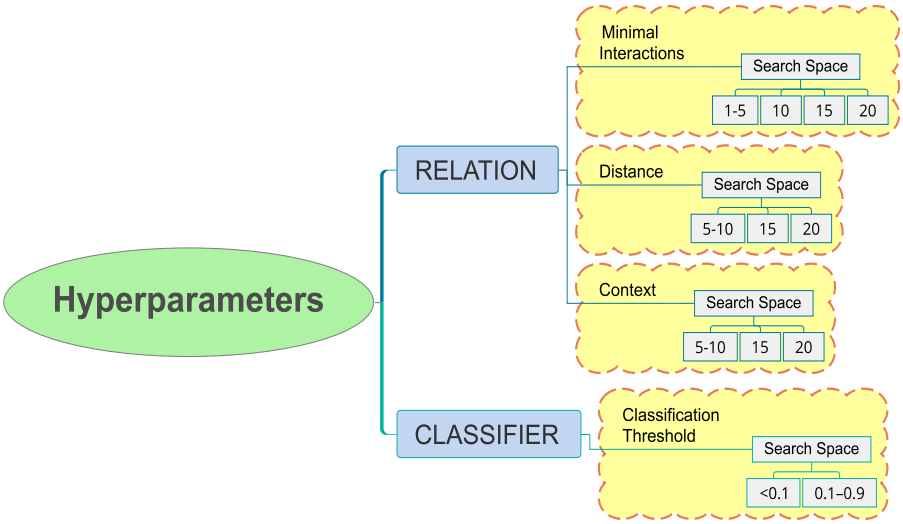


Figure 6.4: Parameter search space for each of the components.

	Dens.	Diam.	Triad.clos.	Avg.deg.	Betw.	F_1
A Clash of Kings	0.11	3	0.1	1.94	0.04	3
A Clockwork Orange	0.79	1.33	0.16	1.21	0.11	0
A Farewell to Arms	0.67	2	0	1.33	0.33	0
A Game of Thrones	0.22	3	0.36	3.75	0.06	6
A Storm of Swords	0.34	3.6	0.44	3.60	0.09	5
A Tale of Two Cities	0.5	3	0	1.75	0.26	0
Absalom, Absalom!	0.31	2	0.58	1.85	0.02	0
Anna Karenina	0.23	3	0.26	2.91	0.07	24
Cat's Cradle	0.93	1.2	0.8	1.86	0.06	21
Dandelion Wine	0.43	2	0.19	1.81	0.17	20
Divergent	0.29	1.83	0.12	1.34	0.04	0
Emma	0.45	2.25	0.41	2.5	0.12	30
Fahrenheit 451	0.63	2.2	0	1.36	0.33	22
Flowers for Algernon	0.91	1.25	0	1.08	0.08	0
Jane Eyre	0.29	3.75	0.24	2.31	0.10	16
Little Women	0.31	3	0.47	4.32	0.05	10
Lord Jim	0.4	2.8	0.05	1.68	0.24	19
Lord of the Flies	0.48	2.5	0.46	2.16	0.18	42
Mansfield Park	0.33	2.6	0.30	2.50	0.10	19
Never Let Me Go	0.37	3	0.23	1.91	0.20	43
Northanger Abbey	0.34	2.33	0	1.65	0.19	33
Pride and Prejudice	0.39	2	0.37	2.63	0.10	28
The Great Gatsby	0.52	2	0	1.47	0.26	31
The Hunger Games	0.66	1	0	1	0	0
The Lion ...	0.58	2.25	0.65	3.03	0.1	29
The Unbearable ...	0.33	2	0	1	0.08	16
Spearman corr. to F_1	0.00	0.21	0.24	0.35	0.40	

Table 6.7: Statistical properties of predicted networks. Bottom row shows Spearman correlation between network properties and F_1 score. Zero indicates the correlation is orthogonal, positive values imply positive correlation.

	diameter	triad.clos.	avg.dens.	between.
density	-0.71	-0.00	-0.45	0.20
diameter		0.09	0.62	0.12
triad.clos.			0.62	-0.43
avg.dens.				-0.28

Table 6.8: Spearman correlation between different network properties.

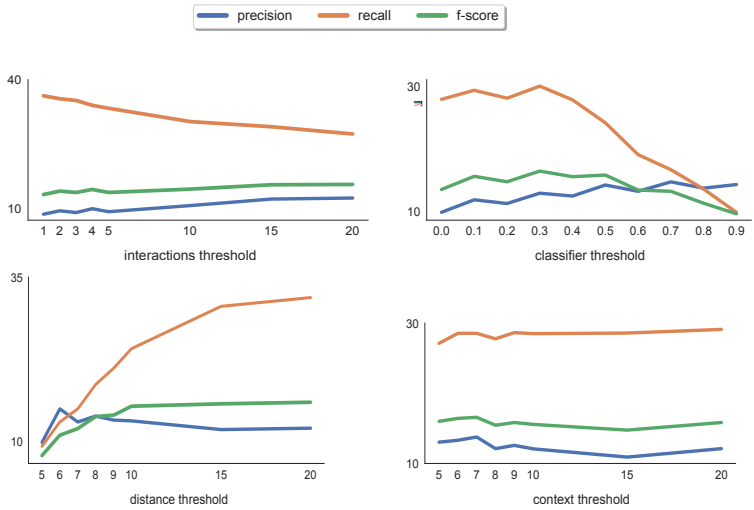


Figure 6.5: Correlation of precision, recall, and F_1 with various parameters.

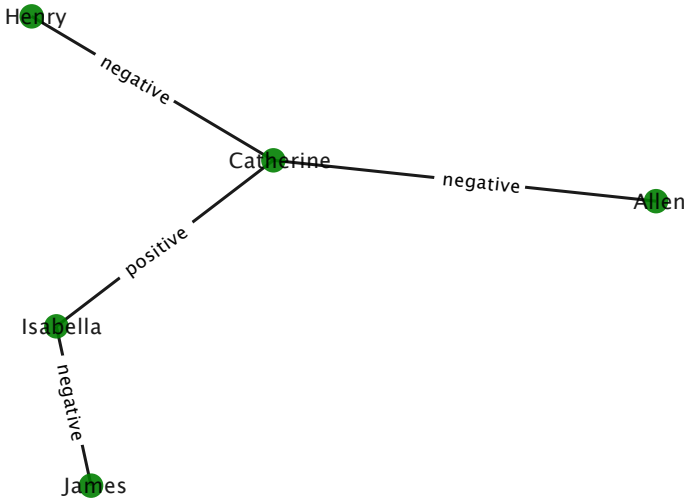


Figure 6.6: The network with binary emotions from *Northanger Abbey* by Jane Austen.

	P	R	F_1	runs	stddev
A Clash of Kings	7	34	11	5	2.92
A Clockwork Orange	0	0	0	6	0
A Farewell to Arms	0	0	0	4	0
A Game of Thrones	10	50	17	6	2.75
A Storm of Swords	14	42	21	5	4.60
A Tale of Two Cities	0	0	0	2	0
Absalom, Absalom!	0	0	0	6	0
Anna Karenina	37	83	51	6	1.57
Cat's Cradle	36	36	36	5	3.86
Dandelion Wine	34	67	45	4	2.38
Divergent	0	1	1	6	2.23
Emma	57	78	65	4	3.11
Fahrenheit 451	46	33	38	5	13.16
Flowers for Algernon	6	4	5	4	8.66
Jane Eyre	30	79	43	4	4.03
Little Women	18	93	31	2	0.5
Lord Jim	37	56	45	5	6.46
Lord of the Flies	52	96	67	2	1.5
Mansfield Park	34	63	43	5	3.77
Never Let Me Go	66	87	75	6	4.67
Northanger Abbey	64	64	63	6	4.76
Pride and Prejudice	47	76	58	5	1.72
The Great Gatsby	56	41	47	6	12.36
The Hunger Games	0	0	0	5	0
The Lion, The Witch ...	59	79	67	4	2.27
The Unbearable Lightness ...	86	51	63	5	4.96
AVERAGE	30	46	34	4.7	
stddev	25.55	33.38	25.81		

Table 6.9: The final pipeline performance achieved via 10-fold cross-validation with binary emotion casting. Each book within each test fold was classified n times, with n ranging from 2 to 6 (due to ties in metaparameter optimization), hence *steddev* for each individual book is given.

Due to small number of samples (26 books used in the pipeline experiments), the classification is done in a leave-one-out fashion using multilayer perceptron. To make the dataset more balanced, I merge science fiction and fantasy into a single class due to only four instances of fantasy books and six instance of science fiction books in the dataset. In the end, there are 16 instances of general fiction, and 10 instances of sci-fi/fantasy. Using fine-grained genre labels (see Table 6.1 on page 179) is not feasible due to several singleton classes (*e.g.*, satire, autobiography, and love story) and very small number of samples.

6.8.3 Results

Table 6.10 on page 191 shows the results of the classification experiment. It can be seen that classification with statistical network features only shows the lowest performance with macro F_1 score of 53%. The classification with combined network and emotion features achieves significantly higher ($p < 0.05$)[10] macro F_1 of 68%. Finally, classification with Emo n features (column E.n) achieves significantly highest macro F_1 of 73%. All the of models except "Net." show significant improvement over the most frequent class baseline (macro and micro F_1 of 38% and 61% respectively).

The number of characters whose emotions are used as features influences F_1. Figure 6.7 on page 192 shows this rela-

[10] Calculated with Wilcoxon singed-ranked test.

	Net.			Net.+Emo			E. n (=20)			E.% (=50)		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
general	64	56	60	79	69	73	**85**	**69**	**76**	**81**	**81**	**81**
sci-fi/fantasy	42	50	45	58	70	64	**62**	**80**	**73**	**70**	**70**	**70**
micro avg F_1	54	54	54	69	69	69	**73**	**73**	**73**	**77**	**77**	**77**
macro avg F_1	53	53	53	68	69	68	**73**	**74**	**73**	**76**	**76**	**76**

Table 6.10: Results of leave-one-out genre classification using network features (Net.), combined network and emotion features (Net.+Emo), and emotion features only (Emo. n and Emo. %). Emo. n features take top n characters sorted by degree in terms of counts (*e.g.*, top 5, top 10 characters). % features take top characters sorted by degree in terms of percentage (*e.g.*, top 20%, top 50% of characters). Bold indicates significantly higher results.

tionship. One can see that the best results are achieved with 20 characters, second best with 10 characters, and third best with all characters taken into account. However, due to the fact that the number of characters changes from book to book, it could happen that for some books 10 and 20 characters is the maximum possible number of characters. Therefore, I repeat the experiment by adapting Emo n features to take into account only *top percentage* of most important characters. For example, instead of taking top n most important characters from each book, the model will take top n% of most important

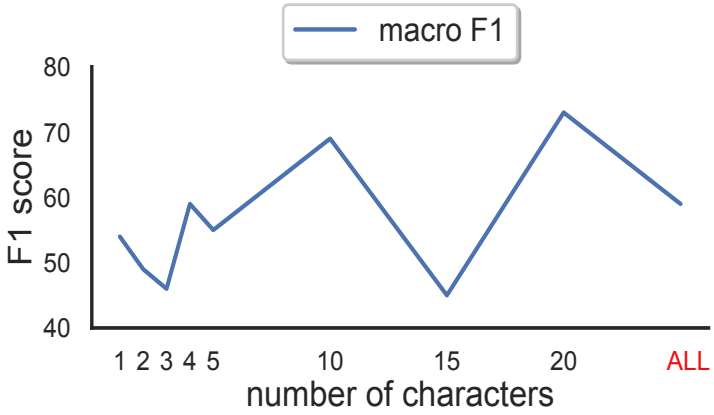


Figure 6.7: Influence of the number of characters whose emotions are considered as features. For example, 20 indicates that emotions of top 20 most important characters are used as features in the classification.

characters (top 30%, top 50%, etc.). This will allow to select features in a more balanced way.

Column E% in Table 6.10 shows that classification with Emo % features achieves significantly higher performance both over network and Emo n features and the best macro F_1 is achieved when 50% of most important characters are taken into account (Figure 6.8 on page 193).

I visualize the differences in emotions between two genres on

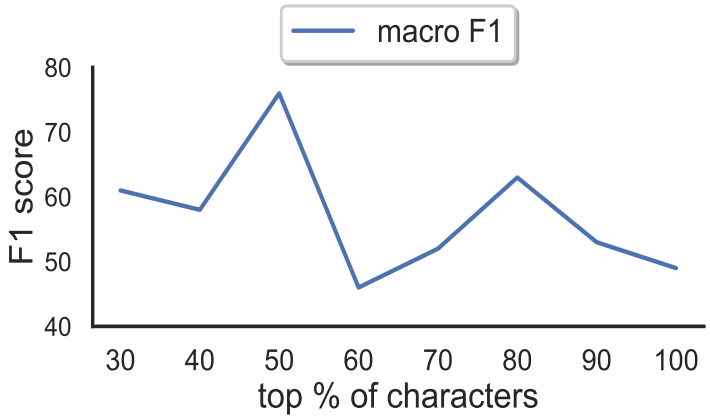


Figure 6.8: Influence of the number of characters – in terms of percentage – whose emotions are considered as features. For example, 50 indicates that emotions of top 50% of most important characters are used as features in the classification.

Figure 6.9 on page 197. It can be seen that there are noticeable differences in emotions associated with the most important characters. Particularly, there is more *anger* and *sadness* in sci-fi/fantasy and more *joy*,*fear*, and *trust* in general fiction. These differences are high enough for a classifier to distinguish between two genres.

6.8.4 Discussion

The results of this experiment show that a multilayer perceptron is capable of distinguishing between two classes of literature – sci-fi/fantasy and general fiction – with maximal macro F_1 of 76%. In particular, general fiction is recognized with F_1 of 81% and sci-fi/fantasy with F_1 of 70%. The best results are achieved with Emo % features that are based on the emotions associated with top 50% of most important characters, whose importance is calculated using the degree.

I conclude that the networks predicted by my model *capture and retain genre information.* However, this information is only useful for genre classification when emotion is taken into account, as evidenced by better performance of all models that consider the emotions of most important characters compared to network-based model. In other words, I find evidence that *emotion features based on the predicted network structure are a good discriminator of genre* in the context of the dataset size and number of predicted classes. This in turn suggests that my pipeline model, in fact, *captures important information* regarding the books I am working with.

6.9 Discussion and Conclusion

In this section, I presented a pipeline that transfers my work on emotion relationships classification to a more realistic scenario,

where character names are not given by default. The goal of this pipeline is to estimate how feasible the task is, and to answer the question raised in the beginning of this section: How to manipulate a precision-recall tradeoff of the social network generation pipeline that uses NER and emotion recognition components?

The analysis of pipeline performance shows that there is no a single set of parameters that are optimal for an arbitrary book. A complex interaction between different pipeline components leads to different optimal parameters on different books. For example, for some books, the classification benefits from less restrictive parameters, such as big distance – in terms of tokens – between characters, but the same parameter may harm the classification of another book. Another bottleneck is error propagation. Mistakes made by the model during a pair assignment step, for example, lead to false positives which harm the precision. On the one hand, this is a downside of the model. On the other hand, I argued that recall should carry more importance for this task, as finding as many pair as possible at this step seems preferable to finding fewer accurately classified pairs. Changing the classifier threshold is one way to adjust model sensitivity: lowering the threshold value allows the model to consider as many emotion-character pairs as possible at the cost of accepting more false positives. At the same time, evaluation of pipeline performance with binary

mapping shows that, at least on polarity level, the results are much more positive (max. F_1=75%), showing that there is a potential for model application in different scenarios.

An experiment on genre classification confirmed that the networks from the books predicted by the pipeline capture genre information, especially when emotion is taken into account along as a structural network property. I believe that it has research potential that could be further addressed in future work.

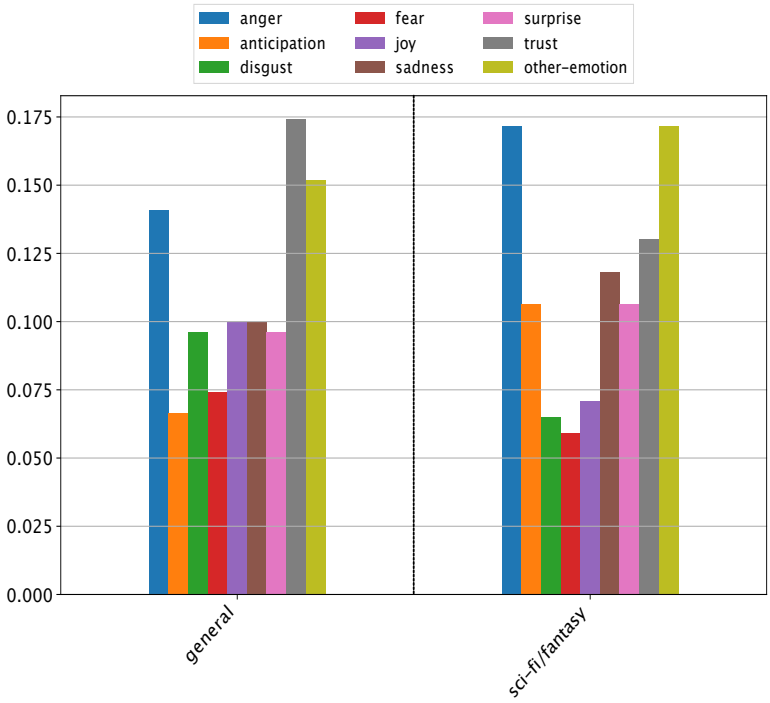


Figure 6.9: Distribution of emotions in genre.

7 Conclusion and Future Work

In my thesis, I focused on several aspects of computational emotion analysis in literature. Namely, I focused on emotions in the context of literary genres, analyzed the linguistic structure of emotions, and modeled the emotional relationships of literary characters.

I started my thesis by testing the usefulness of emotion dictionaries in constructing emotion features for the classification of genres. As it turned out, a "good" emotion dictionary can capture valuable information from a literary text. When used as features in a time-series classifier, the emotional arcs constructed with the dictionary lead to good performance. This finding suggests that 1) emotion is a discriminative feature of a literary text, and 2) emotion dictionaries can be used in the task of emotion analysis. However, there are certain limitations. As analysis shows, genres are not always uniform concerning their prototypical development. Hence, accounting for outliers is an issue that one shall consider before the classifica-

tion. At the same time, as I have established later in my work, the authors use not only single words to describe an emotion, and almost a quarter of all emotions[1] are expressed with two or more words. This may imply that using dictionaries for emotion cues detection will not always work, as emotion dictionaries typically do not include phrases. One way to tackle this problem would be relying on semantic similarity between words in the dictionary and words and phrases of the analyzed text. I propose that future work take these considerations into account.

Additionally, the emotional arcs capture the emotion from the text on a high level, without differentiating between who is the experiencer of the emotion. Therefore, for future work, I propose to build a model of emotional arc construction that creates the arcs on a character, rather than general text level. That is, each character has an associated emotional arc. Characters could further be categorized into the main character and supporting roles, as well as protagonist and antagonist. With this approach, the distinction in emotion plot development could be made separately for different characters, in terms of the role they play in a book, and hence bring more value to the analysis.

Next, I have presented the results of my work on the annotation and classification of the emotion structure. I showed that

[1] In my annotated corpus.

both these tasks are difficult, especially when it comes to predicting emotion *causes*, as they are mostly events and hence are difficult to learn. Similarly for the annotation, the inter-annotator agreement shows that the same category was most difficult for the annotators, especially when it came to an agreement on the exact spans of the respective categories. The main bottleneck in my structured emotion annotation was related to the *relative freedom that annotators had when working with text*. Specifically, I did not impose any linguistic constraints on the annotators during the annotation, although this could have had a positive impact on the quality of the annotations. I suggest that future work that this into consideration. For example, providing the annotators with a syntactic tree for each of the sentences could help them better to see the connections between different parts of the sentence. At the same time, one can introduce further constraints for *cause* annotation, *e.g.*, that it should not cross the borders of a constituent. An even more extreme alternative would be to ask the annotators to annotate the head of a phrase as a *cause* or *target* of an emotion. Such annotation will require a significant amount of preprocessing and preparatory steps but may have lead to better agreement scores. At the same time, future work should take a closer look at the methods used for the classification of emotion structure. A potential research direction is to make use of joint learning of emotion cues and causes similarly to

Xia and Ding (2019), but optimized for literature.

Lastly, in this work, I argued that character relationships are one of the ways emotion manifests itself in a literary text. Therefore, its modeling may provide an interesting angle of analysis for literary scholars. I showed that it is a feasible task benefiting from recurrent neural network architecture with positional indicators. At the same time, we have seen that a real-world application of the model to an arbitrary unseen text is an extremely challenging task. This is mainly explained by the complex interdependency of different pipeline components, from NER to emotion classification, each of which has a set of parameters. These considerations make the real-world application of the pipeline in the immediate future unlikely. At the same time, the results of a small-scale genre experiment show that the networks predicted by the pipeline capture genre information, especially when emotion is taken into account as a structural network property. This, again, suggests that emotion may be a good discriminative feature of a literary text. However, more research is needed with regard to emotion relationship classification in the context of the proposed pipeline. Specifically, more work is needed to solve the precision/recall trade-off bottleneck associated with the pipeline. For future work in this direction, I propose to focus on pair assignment heuristics. A particularly interesting research direction is how characters interact with each other independently of token dis-

tances between them. Indeed, if two characters are mentioned in the same sentence, it may mean that there is an interaction between them, but that cannot always be the case. Therefore, understanding what a character interaction is–irrespectively of the textual measures–is an endeavor that can result in a better understanding of how to model character relationships.

In this work, I only scratched the surface of possible research directions in this area. Nonetheless, I hope I was able to show that computational emotion analysis from literary text is an exciting research area. At the same time, it is a non-trivial task that is often carried out with certain simplifications and not-necessarily-universal generalizations. In my thesis, I adopted a discrete view of emotions, as they were viewed by Paul Ekman and Robert Plutchik. However, modern emotion studies do not focus only on these emotional representations. Such theories as cognitive appraisal theories of Ortony et al. (1990) and Scherer (2005), or a constructionist approach of Russell (2003) and Barrett (2017) offer different, more complex, views of emotions. When applied to a literary text, they could bring a fresh and insightful perspective on emotions in literature, as well as prove or disprove some of the findings I made in my work. Whatever the case, I leave these ideas for future work.

Bibliography

Abdul-Mageed, M. and Ungar, L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Ackerman, A. and Puglisi, B. (2012). *The Emotion Thesaurus: A Writer's Guide to Character Expression*. JADD Publishing.

Agarwal, A., Kotalwar, A., and Rambow, O. (2013). Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1202–1208, Nagoya, Japan. Asian Federation of Natural Language Processing.

Agrawal, A. and An, A. (2012). Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent*

Technology-Volume 01, pages 346–353, Macau, China. IEEE Computer Society.

Agrawal, A., An, A., and Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Alm, C. O. and Sproat, R. (2005). Emotional sequencing and development in fairy tales. In Tao, J., Tan, T., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction*, pages 668–674, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In Matoušek, V. and Mautner, P., editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anderson, C. W. and McMaster, G. E. (1982). Computer assisted modeling of affective tone in written documents. *Computers and the Humanities*, 16(1):1–9.

Anderson, C. W. and McMaster, G. E. (1986). Modeling emotional tone in stories using tension levels and categorical states. *Computers and the Humanities*, 20(1):3–9.

Anderson, C. W. and McMaster, G. E. (1993). Emotional tone in Peter Rabbit before and after simplification. *Empirical Studies of the Arts*, 11(2):177–185.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, page 86–90, USA. Association for Computational Linguistics.

Baker, H. T. (1927). Literature and the emotions. *The English Journal*, 16(10):772–777.

Bal, M. P. and Veltkamp, M. (2013). How does fiction reading influence empathy? an experimental investigation on the role of emotional transportation. *PLOS ONE*, 8(1):1–12.

Baloglu, S. and Love, C. (2005). A cognitive–affective positioning analysis of convention cities: An extension of the

circumplex model of affect. *Tourism Analysis*, 9(4):299–308.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, 12(4):579–599.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, New York.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

Barros, L., Rodriguez, P., and Ortigosa, A. (2013). Automatic classification of literature pieces by emotion detection: A study on quevedo's poetry. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 141–146, Geneva, Switzerland. IEEE.

Barth, F., Kim, E., Murr, S., and Klinger, R. (2018). A reporting tool for relational visualization and analysis of character mentions in literature. In *Book of Abstracts – Digital Humanities im deutschsprachigen Raum*, Cologne, Germany.

Barton, J. (1996). Interpreting character emotions for literature comprehension. *Journal of Adolescent & Adult Literacy*, 40(1):22–28.

Benikova, D., Biemann, C., Kisselew, M., and Pado, S. (2014). GermEval 2014 Named Entity Recognition Shared

Task: Companion Paper. In *Workshop Proceedings of the 12th edition of the KONVENS conference*, pages 104–112, Hildesheim, Germany.

Bentley, R. A., Acerbi, A., Ormerod, P., and Lampos, V. (2014). Books average previous decade of economic misery. *PLOS ONE*, 9(1):1–7.

Bestgen, Y. (1994). Can emotional valence in stories be determined from words? *Cognition & Emotion*, 8(1):21–36.

Bird, M. (2016). *The Secrets of Story: Innovative Tools for Perfecting Your Fiction and Captivating Readers*. Writer's Digest Books, Blue Ash, Ohio.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boot, P., Zijlstra, H., and Geenen, R. (2017). The Dutch translation of the linguistic inquiry and word count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1):65–76.

Booth, P. (2012). The television social network: Exploring TV characters. *Communication Studies*, 63(3):309–327.

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, Barcelona, Spain. ACM.

Bostan, L. A. M., Kim, E., and Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resources Association (ELRA).

Boyles, N. (2012). Closing in on close reading. In Scherer, M., editor, *On Developing Readers: Readings from Educational Leadership, EL Essentials*, chapter 9, pages 89–99. ASCD.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bruggmann, A. and Fabrikant, S. I. (2014). Spatializing a digital text archive about history. In Janowicz, K., Adams, B., McKenzie, G., and Kauppinen, T., editors, *GIO 2014 Geographic Information Observatories co-located with GIScience 2014: Eight International Conference on Geographic Information Science*, CEUR Workshop Proceedings, pages 6–14, Vienna, Austria. CEUR-WS.

Bryant, J. and Zillmann, D. (1984). Using television to alle-

viate boredom and stress: Selective exposure as a function of induced excitational states. *Journal of Broadcasting & Electronic Media*, 28(1):1–20.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Buechel, S., Hellrich, J., and Hahn, U. (2016). Feelings from the Past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 54–61, Osaka, Japan. The COLING 2016 Organizing Committee.

Buechel, S., Hellrich, J., and Hahn, U. (2017). The course of emotion in three centuries of german text–a methodological framework. In *Digital Humanities 2017: Conference Abstracts*, Montreal, Canada.

Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

Cambria, E., Livingstone, A., and Hussain, A. (2012). The hourglass of emotions. *Cognitive behavioural systems*, 7403:144–157.

Chaturvedi, S., Srivastava, S., Daumé, III, H., and Dyer, C.

(2016). Modeling evolving relationships between characters in literary novels. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2704–2710, Phoenix, Arizona.

Chen, A. T., Yoon, A., and Shaw, R. (2012). People, places and emotions: Visually representing historical context in oral testimonies. In *Proceedings of the Third Workshop on Computational Models of Narrative, Istanbul, Turkey*, pages 26–27.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the Deep Learning and Representation Learning Workshop at NIPS 2014*, Montreal.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2015). Gated feedback recurrent neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2067–2075, Lille, France.

Collobert, R. and Bengio, S. (2004). Links between perceptrons, MLPs and SVMs. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML, Banff, Alberta, Canada.

Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised

learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.

Crozier-De Rosa, S. (2010). Popular fiction and the 'emotional turn': The case of women in late victorian britain. *History Compass*, 8(12):1340–1351.

Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry*, 45(3):601–639.

Day, M. (2019). Amazon is working on a device that can read human emotions. *Bloomberg L.P.* retrieved Aug. 12, 2019 from Bloomberg database.

Djikic, M., Oatley, K., and Moldoveanu, M. C. (2013). Reading other minds: Effects of literature on empathy. *Scientific Study of Literature*, 3(1):28–47.

Djikic, M., Oatley, K., Zoeterman, S., and Peterson, J. B. (2009). On being moved by art: How reading fiction transforms the self. *Creativity Research Journal*, 21(1):24–29.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 6(12):1–1.

Eder, J., Jannidis, F., and Schneider, R. (2010). *Characters in fictional worlds: Understanding imaginary beings in literature, film, and other media*, volume 3, chapter Character in Fictional Worlds: An Introduction, pages 3–64. Walter de Gruyter.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Egloff, M., Lieto, A., and Picca, D. (2018). An ontological model for inferring psychological profiles and narrative roles of characters. In *Digital Humanities 2018: Conference Abstracts*, Mexico City, Mexico.

Ekman, P. (1993). Facial expression and emotion. *American psychologist*, 48(4):384.

Ekman, P., Sorenson, E. R., Friesen, W. V., et al. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88.

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644, Avignon, France. Association for Computational Linguistics.

Elsner, M. (2015). Abstract representations of plot structure. *LiLT (Linguistic Issues in Language Technology)*, 12:1–29.

Elson, D., Dames, N., and McKeown, K. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In

Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pages 417–422, Genoa, Italy. European Language Resources Association (ELRA).

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Fillmore, C. J., Petruck, M. R., Ruppenhofer, J., and Wright, A. (2003). Framenet in action: The case of attaching. *International Journal of Lexicography*, 16(3):297–332.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Fish, S. (1970). Literature in the reader: Affective stylistics. *New Literary History*, 2(1):123–162.

Francis, N. W. and Kucera, H. (1979). Brown corpus manual. Received online at `http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM`.

Freytag, G. (1863). *Die Technik des Dramas*. Hirzel, Leipzig, Germany.

Frow, J. (2015). *Genre*. Routledge, New York.

Gao, J., Jockers, M. L., Laudun, J., and Tangherlini, T. (2016). A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4, Duke University, Durham, NC, USA.

Gao, K., Xu, H., and Wang, J. (2014). Emotion classification based on structured information. In *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, pages 1–6, Tsinghua University, Beijing, P.R. China.

Gendron, M. and Feldman Barrett, L. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review*, 1(4):316–339.

Gendron, M., Roberson, D., van der Vyver, J. M., and Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Ghazi, D., Inkpen, D., and Szpakowicz, S. (2010). Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Ap-*

proaches to Analysis and Generation of Emotion in Text, pages 140–146, Los Angeles, CA. Association for Computational Linguistics.

Ghazi, D., Inkpen, D., and Szpakowicz, S. (2015). Detecting emotion stimuli in emotion-bearing sentences. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 152–165, Cham. Springer International Publishing.

Goethe, J. W. v. (1774). *Die Leiden des jungen Werthers*.

Haider, T., Eger, S., Kim, E., Klinger, R., and Menninghaus, W. (2020). PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resources Association (ELRA).

Heise, D. R. (1965). Semantic differential profiles for 1,000 most frequent english words. *Psychological Monographs: General and Applied*, 79(8):1.

Henny-Krahmer, U. E. G. (2018). Exploration of sentiments and genre in spanish american novels. In *Digital Humanities 2018: Conference Abstracts*, Mexico City, Mexico.

Heuser, R., Moretti, F., and Steiner, E. (2016). The emotions

of London. Technical report, Stanford University. Pamphlets of the Stanford Literary Lab.

Hinchman, K. A. and Moore, D. W. (2013). Close reading: A cautionary interpretation. *Journal of Adolescent & Adult Literacy*, 56(6):441–450.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial intelligence*, 40(1):185–234.

Hogan, P. C. (2010). Fictions and feelings: On the place of literature in the study of emotion. *Emotion Review*, 2(2):184–195.

Honnibal, M. (2013). A Good Part-of-Speech Tagger in about 200 Lines of Python. Online: `https://explosion.ai/blog/part-of-speech-pos-tagger-in-python`.

Hoover, D. L., Culpeper, J., and O'Halloran, K. (2014). *Digital literary studies: Corpus approaches to poetry, prose, and drama*, volume 16. Routledge.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hugo, V. (1885). *Les Misérables*. Project Gutenberg: `http://www.gutenberg.org/ebooks/135`.

Ingermanson, R. and Economy, P. (2009). *Writing fiction for dummies*. John Wiley & Sons, Indianapolis, IN.

Jafari, S., Sprott, J. C., and Golpayegani, S. M. R. H. (2016). Layla and Majnun: a complex love story. *Nonlinear Dynam-*

ics, 83(1):615–622.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In *EuroVis (STARs)*, pages 83–103, Cagliari, Italy.

Jannidis, F. (2008). *Figur und Person: Beitrag zu einer historischen Narratologie*, volume 3. Walter de Gruyter.

Jhavar, H. and Mirza, P. (2018). EMOFIEL: Mapping emotions of relationships in a story. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 243–246, Lyon, France. International World Wide Web Conferences Steering Committee.

Jockers, M. L. and Underwood, T. (2016). Text-mining the humanities. In *A New Companion to Digital Humanities*, pages 291–306. Wiley Online Library.

Johnson-Laird, P. N. and Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and emotion*, 3(2):81–123.

Johnson-Laird, P. N. and Oatley, K. (2016). *Handbook of emotions*, chapter Emotions in Music, Literature, and Film, pages 82–97. Guilford Publications.

Joyce, J. (1914). *Dubliners*. Grant Richards, London.

Kakkonen, T. and Galić Kakkonen, G. (2011). SentiProfiler: Creating comparable visual profiles of sentimental content in texts. In *Proceedings of the Workshop on Language Tech-*

nologies for Digital Humanities and Cultural Heritage, pages 62–69, Hissar, Bulgaria. Association for Computational Linguistics.

Kim, E. and Klinger, R. (2018). Who feels what and why? Annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kim, E. and Klinger, R. (2019a). An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Kim, E. and Klinger, R. (2019b). Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

Kim, E. and Klinger, R. (2019c). A survey on sentiment and emotion analysis for computational literary studies. *ZfDG*, 4.

Kim, E., Padó, S., and Klinger, R. (2017a). Investigating the

relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Kim, E., Padó, S., and Klinger, R. (2017b). Prototypical Emotion Developments in Literary Genres. In *Digital Humanities 2017: Conference Abstracts*, pages 288–291, Montréal, Canada. McGill University and Université de Montréal.

Kim, S., Bak, J., and Oh, A. H. (2012). Do You Feel What I Feel? Social Aspects of Emotions in Twitter Conversations. In *International Conference on Web and Social Media*, pages 495–498, Dublin, Ireland.

Klinger, R., de Clercq, O., Mohammad, S. M., and Balahur, A. (2018). IEST: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium. Association for Computational Linguistics.

Klinger, R., Suliya, S. S., and Reiter, N. (2016). Automatic Emotion Detection for Quantitative Literary Studies – A case study based on Franz Kafka's "Das Schloss" and "Amerika". In *Digital Humanities 2016: Conference Abstracts*, Kraków, Poland.

Koolen, C. (2018). Women's books versus books by women.

In *Digital Humanities 2018: Conference Abstracts*, Mexico City, Mexico.

Köper, M. and im Walde, S. S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia. European Language Resources Association (ELRA).

Köper, M., Kim, E., and Klinger, R. (2017). IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark. Association for Computational Linguistics.

Kraicer, E. and Piper, A. (2019). Social characters: The hierarchy of gender in contemporary english-language fiction.

Kuivalainen, P. (2009). Emotions in narrative: A linguistic study of Katherine Mansfield's short fiction. *The Electronic Journal of the Department of English at the University of Helsinki*, 5.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Inter-*

national Conference on Machine Learning, pages 282–289, Williamstown, MA.

Lanham, R. A. (1989). The electronic word: Literary study and the digital revolution. *New Literary History*, 20(2):265–290.

Larsen, R. J. and Diener, E. (1992). Promises and problems with the circumplex model of emotion. *Review of personality and social psychology*, 13:25–29.

Leemans, I., van der Zwaan, J. M., Maks, I., Kuijpers, E., and Steenbergh, K. (2017). Mining embodied emotions: a comparative analysis of sentiment and emotion in dutch texts, 1600–1800. *Digital Humanities Quaterly*, 11(4).

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Liu, B. (2015). *Sentiment Analysis*. Cambridge University Press.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Longo, M. (2020). *Emotions through Literature*. Routledge.

Mar, R. A., Oatley, K., Djikic, M., and Mullin, J. (2011). Emotion and narrative fiction: Interactive influences before, during, and after reading. *Cognition & emotion*, 25(5):818–833.

Marchetti, A., Sprugnoli, R., and Tonelli, S. (2014). Sentiment

analysis for the humanities: the case of historical texts. In *Digital Humanities 2014: Conference Abstracts. University of Lausanne (UNIL) & Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*, pages 254–257.

Marvel, S. A., Kleinberg, J., Kleinberg, R. D., and Strogatz, S. H. (2011). Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences*, 108(5):1771–1776.

Mayer, J. D., Roberts, R. D., and Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, 59(1):507–536.

Miller, H. J. (2014). *Exploring Text and Emotions*, chapter Text; Action; Space; Emotion in Conrad's Nostromo, pages 91–117. Aarhus University Press.

Miralana (2015). What's so strange about a down-home family romance? https://archiveofourown.org/works/5474927.

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Mohammad, S. and Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity,*

Sentiment and Social Media Analysis, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad, S. M. (2012a). # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Mohammad, S. M. (2012b). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

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

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Morin, O. and Acerbi, A. (2017). Birth of the cool: a two-centuries decline in emotional expression in anglophone fiction. *Cognition and Emotion*, 31(8):1663–1675.

Nalisnick, E. T. and Baird, H. S. (2013a). Character-to-

character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, Sofia, Bulgaria. Association for Computational Linguistics.

Nalisnick, E. T. and Baird, H. S. (2013b). Extracting sentiment networks from shakespeare's plays. In *Proceedings of the 12th International Conference on Document Analysis and Recognition*, pages 758–762, Washington, DC. IEEE.

Neill, A. (1991). Fear, fiction and make-believe. *The Journal of Aesthetics and Art Criticism*, 49(1):47–56.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., and Tseng, B. L., editors, *Third International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Nielsen, F. Å. (2011). Afinn. `https://github.com/fnielsen/afinn`.

Oliver, M. B. (1993). Exploring the paradox of the enjoyment of sad films. *Human Communication Research*, 19(3):315–342.

Ortony, A., Clore, G. L., and Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Palmer, A. (2004). *Fictional minds*. University of Nebraska

Press.

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Patti, V., Bertola, F., and Lieto, A. (2015). Arsemotica for arsmeteo. org: Emotion-driven exploration of online art collections. In *The Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*, pages 288–293, Hollywood, Florida. Association for the Advancement of Artificial Intelligence.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of LIWC2007. `http://www.liwc.net/LIWC2007LanguageManual.pdf`.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Phelan, J. (1989). *Reading people, reading plots: Character, progression, and the interpretation of narrative*. University of Chicago Press.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on Global Word-*

Net, Mysore, India.

Picard, R. W. (2000). *Affective computing*. MIT press, Cambridge, MA.

Plutchik, R. (1991). *The Emotions*. University Press of America, Lanham.

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Reed, E. (2018). Measured unrest in the poetry of the black arts movement. In *Digital Humanities 2018: Conference Abstracts*, Mexico City, Mexico.

Richard, J. (2009). Moral relativists gone wild. *Mind*, 118.

Richins, M. L. (1997). Measuring emotions in the consumption experience. *Journal of consumer research*, 24(2):127–146.

Rinaldi, S., Landi, P., and ROSSA, F. D. (2013). Small discoveries can have great consequences in love affairs: the case of beauty and the beast. *International Journal of Bifurcation and Chaos*, 23(11):1330038.

Robinson, J. (2005). *Deeper than reason: Emotion and its role in literature, music, and art*. Oxford University Press, New York.

Ross, C. S. (1999). Finding without seeking: the information encounter in the context of reading for pleasure. *Information Processing & Management*, 35(6):783–799.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102.

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

Russell, J. A., Bachorowski, J.-A., and Fernández-Dols, J.-M. (2003). Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1):329–349.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805–819.

Russo, I., Caselli, T., Rubino, F., Boldrini, E., and Martínez-Barco, P. (2011). EMOCause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 153–160, Portland, Oregon. Association for Computational Linguistics.

Sætre, L., Lombardo, P., and Zanetta, J. (2014). *Exploring Text and Emotions*, chapter Text and Emotions, pages 9–

26. Aarhus University Press.

Samothrakis, S. and Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS one*, 10(11):e0141922.

Samur, D., Tops, M., and Koole, S. L. (2018). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition and Emotion*, 32:130–144.

Scarantino, A. (2016). The Philosophy of Emotions and Its Impact on Affective Sciences. In Lewis, M., Haviland-Jones, J. M., and Barrett, L. F., editors, *Handbook of Emotions*, chapter 1, pages 3–49. Guilford Publications, New York.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

Sharar, S. R., Alamdari, A., Hoffer, C., Hoffman, H. G., Jensen, M. P., and Patterson, D. R. (2016). Circumplex

model of affect: a measure of pleasure and arousal during virtual reality distraction analgesia. *Games for health journal*, 5(3):197–202.

Sharoff, S., Wu, Z., and Markert, K. (2010). The web library of babel: evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3063–3070, Valletta, Malta. European Language Resources Association (ELRA).

Smith, H. and Schneider, A. (2009). Critiquing models of emotions. *Sociological Methods & Research*, 37(4):560–589.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3 – 14.

Sprugnoli, R., Tonelli, S., Marchetti, A., and Moretti, G. (2016). Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4):762–772.

Srivastava, S., Chaturvedi, S., and Mitchell, T. (2016). Inferring interpersonal relations in narrative summaries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2807–2813, Phoenix, Arizona.

Stimson, F. J. (1943). *The King's Men: A Tale of Tomorrow*. Project Gutenberg: `http://www.gutenberg.org/ebooks/18960`.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1968). The general inquirer: A computer approach to content analysis.

American Journal of Sociology, 73(5):634–635.

Strapparava, C. and Valitutti, A. (2004). WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1083–1086, Lisbon, Portugal. European Language Resources Association (ELRA).

Suttles, J. and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 121–136, Berlin, Heidelberg. Springer Berlin Heidelberg.

Taboada, M., Gillies, M. A., and McFetridge, P. (2006). Sentiment classification techniques for tracking literary reputation. In *LREC workshop: towards computational models of literary analysis*, pages 36–43, Genoa, Italy.

Taboada, M., Gillies, M. A., McFetridge, P., and Outtrim, R. (2008). Tracking literary reputation with text analysis tools. In *Meeting of the Society for Digital Humanities*.

Tolstoy, L. (1962). *What is Art?: and Essays on Art*, volume 331. Reprint Services Corp.

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, New York.

Tseng, A., Bansal, R., Liu, J., Gerber, A. J., Goh, S., Posner, J., Colibazzi, T., Algermissen, M., Chiang, I.-C., and Rus-

sell, J. A. (2014). Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. *Journal of autism and developmental disorders*, 44(6):1332–1346.

Underwood, T. (2016). The life cycles of genres. *Journal of Cultural Analytics*.

Van Horn, L. (1997). The characters within us: Readers connect with characters to create meaning and understanding. *Journal of Adolescent & Adult Literacy*, 40(5):342–347.

van Meel, J. M. (1995). Representing emotions in literature and paintings: a comparative analysis. *Poetics*, 23(1-2):159–176.

Vanhoutte, E. (2013). The gates of hell: History and definition of digital| humanities| computing. In Terras, M., Nyhan, J., and Vanhoutte, E., editors, *Defining Digital Humanities: A Reader*, chapter 6, pages 119–150. Ashgate Surrey.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Xia, R. and Ding, Z. (2019). Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Yanal, R. J. (1999). *Paradoxes of emotion and fiction*. Penn

State Press, University Park, PA.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327–343.

Zehe, A., Becker, M., Hettinger, L., Hotho, A., Reger, I., and Jannidis, F. (2016). Prediction of happy endings in German novels based on sentiment information. In *3rd Workshop on Interactions between Data Mining and Natural Language Processing, Riva del Garda, Italy*, pages 9–16.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

Zhuravlev, M., Golovacheva, I., and de Mauny, P. (2014). Mathematical modelling of love affairs between the characters of the pre-masochistic novel. In *2014 Second World*

Conference on Complex Systems (WCCS), pages 396–401.

Zillmann, D., Hezel, R. T., and Medoff, N. J. (1980). The effect of affective states on selective exposure to televised entertainment fare. *Journal of Applied Social Psychology*, 10(4):323–339.