Department of Digital Humanities
2021

The Language of Emotions:
Building and Applying Resources for Computational Approaches to Emotion Detection for English and Beyond

Emily S. Öhman

Doctoral dissertation, to be presented for public discussion with the permission of the Faculty of Arts of the University of Helsinki on the 5th of March, 2021 at 16 o'clock. The defence is open for audience through remote access.

University of Helsinki
Finland

Supervisor

Prof. Jörg Tiedemann, University of Helsinki, Finland
Prof. Timo Honkela, University of Helsinki, Finland
Prof. Tanja Säily, University of Helsinki, Finland

Pre-examiners

Prof. Cecilia Ovesdotter Alm, Rochester Institute of Technology, USA
Prof. Cornelius Puschmann, University of Bremen, Germany

Opponent

Prof. Cecilia Ovesdotter Alm, Rochester Institute of Technology, USA

Custos

Prof. Jörg Tiedemann, University of Helsinki, Finland

Contact information

Department of Digital Humanities
P.O. Box 24 (Unioninkatu 40)
FI-00014 University of Helsinki
Finland

Email address: postmaster@helsinki.fi
URL: http://www.helsinki.fi/
Telephone: +358 9 1911, telefax: +358 9 191 51120

ISBN 978-951-51-7105-4 (printed)
ISBN 978-951-51-7106-1 (PDF)

Helsinki 2021
Unigrafia

The Language of Emotions: Building and Applying Resources for Computational Approaches to Emotion Detection for English and Beyond

Emily S. Öhman

Department of Digital Humanities
P.O. Box 24, FI-00014 University of Helsinki, Finland
emily.ohman@helsinki.fi

PhD Thesis, 2021
Helsinki, February 2021, 174 pages
ISBN 978-951-51-7105-4 (printed)
ISBN 978-951-51-7106-1 (PDF)

Abstract

Emotions have always been central to the human experience: the ancient Greeks had philosophical debates about the nature of emotions and Charles Darwin can be said to have founded the modern theories of emotions with his study *The expression of the emotions in man and animals*. Theories of emotion are still actively researched in many different fields from psychology, cognitive science, and anthropology to computer science.

Sentiment analysis usually refers to the use of computational tools to identify and extract sentiments and emotions from various modalities. In this dissertation I use sentiment analysis in conjunction with natural language processing to identify, quantify, and classify emotions in text. Specifically, emotions are examined in multilingual settings using multidimensional models of emotions.

Plutchik's wheel of emotions and emotional intensities are used to classify emotions in parallel corpora via both lexical methods and supervised machine learning methods.

By analyzing emotional language content in text, the connection between language and emotions can be better understood. I have developed new approaches to create a more equitable natural language processing approach for sentiment analysis, meaning the development and evaluation of massively multilingual annotated datasets, contributing to the provision of tools for under-resourced languages.

This dissertation is comprised of ten articles on related topics in sentiment analysis. In these articles, I discuss lexicon-based methods and the creation of emotion and sentiment lexicons, the creation of datasets for supervised machine learning, the training of models for supervised machine learning, and the evaluation of such models. I also examine the annotation process in relation to creating datasets in depth, including the creation of a light-weight easily deployed annotation platform. As an additional step, I test the different approaches in downstream applications.

These practical applications include the study of political party rhetoric from the perspective of emotion words used and the intensities of those emotion words. I also examine how simple lexicon-based methods can be used to make the study of affect in literature less subjective. Additionally, I attempt to link sentiment analysis with hate speech detection and offensive speech target identification.

The main contribution of this dissertation is in providing tools for sentiment analysis and in demonstrating how these tools can be augmented for use in a wide variety of languages and practical applications at low cost.

Tiivistelmä

Tunteet ovat aina olleet keskeisessä asemassa ihmiskokemuksessa: muinaisilla kreikkalaisilla oli filosofisia keskusteluja tunteiden luonteesta, ja Charles Darwinin voidaan sanoa perustaneen modernit tunneteoriat kirjallaan *Tunteiden ilmaisu ihmisissä ja eläimissä*. Tunneteorioita tutkitaan edelleen aktiivisesti monilla eri aloilla psykologiasta, kognitiotieteestä ja antropologiasta tietojenkäsittelytieteeseen.

Sentimenttianalyysi viittaa yleensä tunteiden tunnistamiseen ja erottamiseen laskennallisin menetelmin eri modaliteeteissä. Tässä väitöskirjassa käytän sentimenttianalyysiä yhdessä kieliteknologian menetelmien kanssa tunteiden tunnistamiseksi, kvantifioimiseksi ja luokittelemiseksi tekstissä. Erityisesti tutkin tunteita ja niiden käyttöä monikielisissä ympäristöissä käyttäen moniulotteisia tunnemalleja.

Hyödynnän Plutchikin tunnepyörää ja emotionaalisia intensiteettejä rinnakkaiskorpuksissa esiintyvien tunteiden luokittelussa sekä leksikaalisten että valvottujen koneoppimismenetelmien avulla. Analysoimalla tekstin emotionaalista kieltä voidaan paremmin ymmärtää kielen ja tunteiden välistä yhteyttä. Olen kehittänyt uusia lähestymistapoja helpottamaan massiivisesti monikielisen annotoidun datan kehittämistä ja arviointia, mikä osaltaan auttaa tarjoamaan työkaluja pieniresurssisille kielille.

Tämä väitöskirja koostuu kymmenestä artikkelista, jotka kytkeytyvät eri tavoin sentimenttianalyysiin. Näissä artikkeleissa käsittelen leksikkopohjaisia menetelmiä, sentimentti- ja emootioleksikkojen kehitystä, datan keräystä koneoppimista varten sekä koneoppimismallien arviointia datan avulla. Lisäksi sovellan näitä lähestymistapoja yhteiskuntatieteissä ja kirjallisuustieteissä. Tutkin myös annotaatioprosessin vaikutusta annotoijien tuottaman datan laatuun Sentimentator–annotaatiotyökalun kehitystyöhon pohjautuen.

Tämän väitöskirjan tärkein kontribuutio on tekstianalyysityökalujen kehittäminen sekä niiden hyödyllisyyden osoittaminen useilla eri kielillä ja erilaisissa käytännön sovelluksissa.

General Terms:
sentiment analysis, emotion detection, annotation

Additional Key Words and Phrases:
annotation projection, emotion classification, system evaluation, low-resource languages, emotion translation, crosslingual resources, transfer learning, digital humanities

Preface

Ever since elementary school I told people I was going to be a researcher when I grew up. I wasn't quite sure if the field would be medical research, astrophysics, or mathematics, but I knew for sure this is what I would do. I used to have a binder with graph paper next to bed and would come up with ideas about space and numbers while lying in bed trying to sleep. I did go to a boarding school for budding mathematicians for high school, and studied engineering and mathematics for a few years before linguistics took over. My love for language and logic was well-suited for the rise of computational linguistics and digital humanities.

Almost by accident, I ended up working in Japan and Australia for nearly a decade before returning to my academic endeavors. I felt immediately at home in the Department of Modern Languages at the University of Helsinki where I worked as a research assistant on the LCD-project under Professor Terttu Nevalainen — a job I doubt I would have been hired for if it was not for a lucky coincidence; my MA thesis supervisor in Sweden was an old student of Terttu's: Professor Mikko Laitinen. Mikko steered me in the direction of diachronic corpus linguistics —exactly what the LCD project needed — and definitely made me a better writer and researcher. I worked on the LCD project for a few years before submitting my dissertation proposal. I learned so much from that project about what research really is, how projects work, and how to publish, and also met Tanja Säily, whose doctoral defense was the first I ever attended and who would later become one of my doctoral supervisors.

So many people have supported me during these last few years. First and foremost I would like to express my gratitude to my amazing supervisors without whom this dissertation would not exist:

Timo Honkela was a truly remarkable person and a true visionary. I am grateful for his support and input up to the final stretches of my doctoral work. Timo was the best guide one could hope to have through the philosophical aspects of the PhD journey. He was always full of fresh ideas and enthusiasm towards new takes on old problems. I am sad he can not be here today, but I hope that my dissertation would have made him proud.

Jörg Tiedemann is everything one could want in a supervisor. He setup regular meetings early on, and prepared me for the realities of the academic world. He helped me both with the details of my work as well as the big picture. He also early on tried to teach me to say “no” to different side-projects, something I am still learning as attested by my **two** post doctoral projects. Jörg always had my back and I am grateful for his invaluable input on the PhD process as a whole and specifically for his help and suggestions for improving my dissertation. I was shocked to say the least when a few weeks before the submission deadline he suggested that I completely change the structure of my dissertation. Somehow I managed to complete the changes, and it is clear to me now, that he saw the core of what I had been trying to say much more clearly than I did and my dissertation is much better off because of this.

Tanja Säily came on board the supervisor team in the later stages to offer a link to Digital Humanities when it became clear that my dissertation would not be a typical language technology PhD. Tanja has a keen eye for detail, which saved me from embarrassing mistakes countless times, but more than that, she was always willing to give me her honest opinion and support.

I would also like to thank the pre-examiners, Professor Cornelius Puschmann and Professor Cecilia Ovesdotter Alm, for their invaluable comments on the first version of my dissertation. Their suggestions helped polished and improve the thesis immensely. An extra thanks go to Professor Ovesdotter Alm for also agreeing to be my opponent at my doctoral defense. In addition, I want to thank Dr. Mathias Creutz and Professor Eetu Mäkelä for agreeing to be members of the grading committee despite their busy schedules.

I would also like to thank the Doctoral Programme in Language Studies (HELSLANG) at the University of Helsinki for believing in my project and granting me a

4-year salaried doctoral position – which was extended due to maternity leave – through which I was able to focus on my doctoral research full-time. Not only was I lucky enough to receive a salary-grant, I also received several travel grants from both the faculty, the doctoral program, and the BAULT (Building and Understanding Language Technology) project, which enabled me to present at several international conferences per year, attend the Lisbon Machine Learning Summer School, and spend three months as a visiting researcher at the University of Tokyo. Furthermore, the Doctoral Student Services, namely the Academic Affairs Officers Jutta Kajander and Anniina Sjöblom, kept me sane during the stressful time before the defense and were immensely helpful with the administrative aspects related to submitting and printing the dissertation. Thank you.

I was lucky to enter the Department of Modern Languages at a time when Digital Humanities was taking over. The Department structure changed and Language Technology became a part of the Department of Digital Humanities as the Department of Modern Languages ceased to exist. My research has always had an interdisciplinary flair and the increased focus on digital humanities served me well enabling me to join interdisciplinary projects where I could use both my computational skills and domain expertise to find new approaches to access hidden knowledge in underexplored dataset.

I am grateful for all the support from the Language Technology team at the University of Helsinki. Special thanks go to Miikka Silfverberg and Robert Östling who were always willing to help with Python issues beyond StackOverflow. I was lucky enough to not only have the support from my own Language Technology group, but also the Digital Humanities and Rajapinta crews. Especially Professors Mikko Tolonen and Eetu Mäkelä, who were always willing to help and come up with solutions when I knocked on their door and have read, advised, and commented on tens of papers and grant applications of mine helping me improve them. Thanks also go to Dr. Salla-Maaria Laaksonen and Dr. Matti Nelimarkka, who provided countless insights into cross-disciplinary computational methods, research ethics, and conversations about progressing inter-disciplinarity at Finnish Universities.

I would also like to thank Professor Kyo Kageura of the University of Tokyo for agreeing to host me for three months during which he was genuinely interested in furthering my academic progress through inviting me to take part in his Bias in Machine

Learning project as well as through our bi-weekly discussions about my research. And in conjunction I would also like to thank Piao Hui, a doctoral student of translation studies under Professor Kageura, who helped on the translation studies aspect of one of my papers and welcomed me at the University of Tokyo Library and Information Science lab.

Many of my papers would not exist without my co-authors and graduate students Kaisla Kajava and Marc Pàmies who contributed to several projects with invaluable insights, code, and detailed knowledge of methods. I feel privileged to have been able to supervise their graduate theses. I am also grateful for my co-authors on the more interdisciplinary projects dealing with applied sentiment analysis, Juha Koljonen, Pertti Ahonen, and Mikko Mattila, as well as the wonderful Riikka Rossi. They all gave me insights into how other fields work and made me a better researcher by widening my perspective.

When you use a researcher's work in most of your own work, and see them hold keynote speeches at conferences, they become something akin to academic rock stars in your mind (or maybe that is just me). To me this rock star is Dr. Saif Mohammad, the creator of the NRC Emotion lexicons, without whom none of my work would exist in its current form. While finishing my dissertation, I sent him a draft version of my lexicon paper, and he immediately volunteered his time to talk about improvements to the paper which were much appreciated. I only regret not contacting him sooner.

I would probably not be here writing the preface to my doctoral dissertation if it was not for my grandfather, who from an early age instilled in me the value of knowledge and provided me with endless books and challenging conversations throughout my childhood and adolescence. Thanks go also to my parents who provided support in my academic endeavors ever since my mom asked my first grade teacher why we never had homework (we did), my aunt Annika, who always made me feel less weird for having read all of Agatha Christie's works by second grade, and my brother who believed in me more than I ever did myself.

Most of all, I would like to thank my family who has supported me throughout this process. Especially, my husband, Jesse, who without hesitation took parental leave so that I could go back to my research and once took a week off work to be with our kids so

that I could attend a conference on the other side of the world with only three days notice. I especially want to thank Jesse for the patience he showed when I was finalizing my dissertation and worked seven days a week, often 12 hours a day. Thank you to my kids who have cheered me on and sat next to me during the COVID-19 lockdown hammering on broken keyboards pretending to be my co-workers. You are the best and it is for you I have had the tenacity to finish this process.

Original Papers

This dissertation consists of an introduction and the following peer-reviewed publications, which are referred to as Papers I–X in the text. These publications are reproduced at the end of the print version of the thesis.

I Emily Öhman. SELF & FEIL: Sentiment and Emotion Lexicons for Finnish. Submitted to *Proceedings of the The 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*, Reykjavik, Iceland, 2021.

II Emily Öhman and Kaisla Kajava. Sentimentator: Gamifying Fine-grained Sentiment Annotation. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, Helsinki, Finland, 2018.

III Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. Creating a Dataset for Multilingual Fine-grained Emotion-detection Using Gamification-based Annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, 2018.

IV Emily Öhman. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, 2020.

V Emily Öhman, Timo Honkela, and Jörg Tiedemann. The Challenges of Multidimensional Sentiment Analysis Across Languages. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, Osaka, Japan, 2016.

VI Kaisla, Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann. Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, 2020.

VII Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference of Computational Linguistics*, Barcelona, Spain, 2020.

VIII Marc Pàmies, Emily Öhman, Kaisla Kajava, and Jörg Tiedemann. LT@Helsinki at SemEval-2020 Task 12: Multilingual or language–specific BERT? In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain, 2020.

IX Juha Koljonen, Emily Öhman, Mikko Mattila, and Pertti Ahonen. Strength and intensity of sentiments and emotions in party manifestos: Finland, 1945 to 2019. Submitted to *Scandinavian Political Studies*

X Emily Öhman and Riikka Rossi. Affect and Emotions in Finnish Literature: Combining Qualitative and Quantitative Approaches. Submitted to *LaTeCH-CLfL 2020: The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities, and Literature*

Three papers by the author have **not been included** in this dissertation. These papers are:

XI Emily Öhman. Teaching Computational Methods to Humanities Students. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, Copenhagen, Denmark, 2019.

XII Terttu Nevalainen, Turo Vartiainen, Tanja Säily, Joonas Kesäniemi, Agatha Dominowska, and Emily Öhman. Language Change Database: A new online resource. In *ICAME journal*, 40(1), pp.77-94, 2016.

XIII Emily Öhman. Case Study: Bank. In Allwood, J., Lindström, N.B., Börjesson, M., Edebäck, C., Myhre, R., Voionmaa, K. and Öhman, E., (eds.) *European intercultural workplace: Sverige*, Göteborgs Universitet, 2007.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The field of sentiment analysis has its roots in the field of public opinion analysis in the early 1900s and text subjectivity analysis research conducted in the 1990s (Mäntylä et al., 2018). The earliest attempts to systematically extract sentiments and opinions from text tended to rely on sentiment lexicons. Advances in computer science in general, and machine learning and natural language processing in particular, were quickly incorporated into sentiment analysis methods too and the field is still rapidly advancing.

According to the definition of Merriam-Webster (Merriam-Webster Online, 2009) *sentiment* is understood as an attitude, thought, or judgment prompted by feeling, a specific view or notion, or an idea colored by emotion. In the field of natural language processing, sentiment analysis typically refers to the automatic detection of such attitudes in texts or other modalities. It is often simplified into a classification task with the polarities of a text sorted into disjoint categories of positive, negative and sometimes also neutral sentiments. In some cases the categories are less disjoint and presented on a sliding valence scale from e.g. -1 to 1. Emotion detection, on the other hand, tries to detect more specific emotions in text such as *happy*, *sad*, and *angry*. These terms are explored in more detail in section 2.1.

1.1 Motivation

As human beings we are curious about our surroundings, and our surroundings usually include other people. We are also curious about what those other people think, how they feel, and especially how their feelings relate to us. With the rise of online platforms where people are free to express their opinions and vent their feelings about various topics, items, goods, events, etc., it is not strange to see a simultaneous rise in interest in automatically determining how people feel about these things.

The interest in opinion mining and sentiment analysis has been going strong for over two decades now and shows only signs of increased interest, with more than 7000 papers written on the topic in the spring of 2017 and 99% of them since 2004 (Mäntylä et al., 2018; Banea et al., 2008). A Google Scholar search[1] on *sentiment analysis* with quotes gives 154,000 hits. The simultaneous spread of online platforms and the sub-

[1] Search conducted on January 15, 2021.

sequent rise of big data means that there is more data to work with than ever before. This data has many practical and scholarly applications. Companies are curious about how their brand is discussed online and what that does for their bottom line and, in some cases, stock value (Bollen et al., 2011), but sentiment analysis can also be used to predict box-office performance of movies (Asur and Huberman, 2010), analyze or predict the outcome of elections (Tumasjan et al., 2010; Budiharto and Meiliana, 2018; Koljonen et al., forthcoming (submitted; Mohammad et al., 2015), predict and analyze disasters (Desai et al., 2020; Ragini et al., 2018; Beigi et al., 2016), and to detect hate speech (Pàmies et al., 2020; Felt and Riloff, 2020; Davidson et al., 2017), to name a few practical applications. Sentiment analysis has also been used to study city-planning (Ali et al., 2017; Flaes et al., 2016; Li et al., 2016) and as a support for literary analysis (Kim and Klinger, 2018; Alm et al., 2005; Mohammad, 2012; Elsner, 2012).

Sentiment analysis has greatly benefited from the progress made in the fields of data mining and machine learning, but the reverse is also true with one researcher calling it "The Unstoppable Rise of Computational Linguistics in Deep Learning" (Henderson, 2020), suggesting that language technology and computational linguistics, which sentiment analysis is a part of, is becoming the focus of most machine learning and deep learning tasks and therefore shaping the development of these fields in return. It seems safe to assume that as deep- and machine learning models and algorithms advance, so will the field of sentiment and emotion analysis.

For a long time, sentiment analysis mainly focused on the binary classification scheme of positive and negative, and almost all publicly available datasets for sentiment analysis were created for the English language (He and McAuley, 2016; Socher et al., 2013; Blitzer et al., 2007; Maas et al., 2011; Go et al., 2009). Binary or ternary (the addition of neutral) sentiment analysis can be useful in many applications, but a more precise division into emotions rather than sentiments allows for the extraction of more detailed insights from the data at hand. A binary or ternary approach can be seen as two-dimensional (moving on one axis of polarity), whereas more fine-grained emotions are termed multi-dimensional in the research presented here.

English is the language which most tools and models are aimed at and optimized for. Because English is also the new lingua franca in academia and academic research, even

those whose native language is not English tend to publish in English. Furthermore, even non-native researchers tend to study the English language and therefore the resources they create are often for English. This means that there is also disproportionate growth in datasets available for English. There are, however, many practical applications where datasets in other languages are needed.

The research presented in this dissertation aims at narrowing the gap between English and other languages from a natural language processing (NLP) point of view in terms of availability of datasets and related tools such as annotation guidelines and platforms. Creating annotated datasets is expensive and time-consuming, and often beyond the scope and budget of projects. This is even more true for low-resource languages that are mainly spoken in economically disadvantaged countries and regions. There are various ways of bootstrapping resources for new languages. Our focus is on annotation projection, specifically from English, as a simple and cheap way of producing quality datasets for languages with fewer linguistic resources.

1.2 Research Questions and Objectives

The studies that this dissertation is comprised of examine different methods to extract emotions from text. Beyond sentiment analysis and emotion detection, the studies provide annotation tools, emotion lexicons, and multi-lingual datasets annotated for emotions. The overarching research questions are:

- What are the advantages and disadvantages of different sentiment analysis and emotion detection methods and how can these be improved upon?
- How well are sentiments and emotions preserved in translation?
- Is cross-lingual sentiment analysis robust enough to be useful for improving sentiment analysis datasets for low-resource languages? Is annotation projection a feasible and reliable method to create datasets for low-resource languages?
- How well do models beyond positive-negative fare in real-world tasks?

These questions are addressed in the papers that comprise this dissertation and analyzed as a whole in the summary and synthesis part of it. The main objective is to examine the usefulness of specific types of approaches, and to evaluate the different systems used in these approaches. Another goal is to explore annotation projection from one language to another in the creation of annotated open datasets for low resource languages.

1.3 Contributions

The main contribution of this thesis is a deeper understanding of fine-grained emotion detection in general, but also specifically multi- and cross-lingual emotion detection across domains. The outcome of this is increased understanding of these topics, but also a practical contribution in the creation of tools and datasets for emotion detection, specifically for non-English languages. Among these tools is the emotion annotation platform *Sentimentator*[2], which offers a simple portable annotation platform that is optimized for emotion annotation but can easily be used for other types of text annotation, and for different granularities. It is lightweight, dockerized, and freely available on GitHub[3].

Using *Sentimentator* we created an annotated dataset for emotion detection we call XED (CROSSlingual Emotion Detection). By using annotation projection and parallel corpora we extended the data to cover multiple languages, with more than 10,000 annotated sentences for 12 languages. These datasets are available on GitHub.

In addition, the Finnish version of the NRC Emotion lexicon (Mohammad, 2012) and the NRC Emotion Intensity Lexicon (Mohammad, 2018b) were augmented and carefully adjusted to reflect better translations. Duplicates and other types of noise were also culled from the lexicon. I call the new lexicons SELF (Sentiment and Emotion Lexicon for Finnish), and FEIL (Finnish Emotion Intensity Lexicon)[4].

Finally, these tools and resources have been evaluated by using them in cross--disciplinary projects including political science, hate-speech detection, and comparative literature.

[2] `http://sentimentator.it.helsinki.fi`

[3] `https://github.com/Helsinki-NLP/sentimentator`

[4] `https://github.com/Helsinki-NLP/SELF-FEIL`

1.4 Structure of the Dissertation

The Introduction chapter of this dissertation presents an overview of content and contributions. The second chapter provides an overview of theoretical research into emotions from a mainly linguistics and cultural point of view. In Chapter 3, I discuss annotation in theory and practice, present an overview of existing datasets, lexicons, and annotation platforms and other tools. In Chapter 4, I discuss the different methods used for sentiment analysis and emotion detection and previous work on the same topics, including downstream applications.

After the first four chapters that set the background, Chapter 5 presents my contributions starting with the improvement of existing lexicons, the creation of annotation tools and an examination of annotator motivation, followed by an investigation of emotion preservation in translation, and finally the emotion dataset, XED, and its multilingual extensions created via annotation projection. Before concluding the chapter I discuss the usefulness of the tools and resources I have created by applying them to practical downstream tasks.

In the concluding chapter I discuss the findings of the preceding chapters and the different research papers, and analyze the results as a whole. The research questions presented in section 1.2 are explicitly answered and the scholarly and practical contributions are discussed.

The final part of this dissertation consists of the papers and articles that are the foundation of my doctoral research.

1.5 Overview of Articles

This section outlines the author's contributions to each article and published paper that the dissertation is comprised of. This dissertation consists of ten articles of which the majority are by multiple authors, as is typical in the fields of language technology, computational linguistics, and computer science. In this section, my responsibilities and contributions are outlined paper by paper. The papers are referenced in the order they were presented earlier, and are grouped by roughly five topics:

1. The creation of emotion lexicons (paper I)

2. Annotation theory and tools (papers II, III, and IV)

3. Sentiment and emotion preservation in translation (papers V and VI)

4. Multilingual emotion datasets (paper VII)

5. Practical applications (papers VIII, IX, and X)

When deciding on the order of authors, and the inclusion of authors, the principles of McNutt et al. (2018) and those of ICMJE[5] were followed. All authors included have contributed to the study conception and/or design. All published articles are under the CC BY 4 license[6] of the original publisher. Unpublished articles are reprinted as manuscripts and might differ slightly from their final published format.

I Emily Öhman. SELF & FEIL: Sentiment and Emotion Lexicons for Finnish. Submitted to *Proceedings of the The 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*, Reykjavik, Iceland, 2021.
This paper introduces the Sentiment and Emotion Lexicon for Finnish and the Finnish Emotion Intensity Lexicon. All work was completed solely by Emily Öhman.

II Emily Öhman and Kaisla Kajava. Sentimentator: Gamifying Fine-grained Sentiment Annotation. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, Helsinki, Finland, 2018.
The concept for this paper came from Emily Öhman. The platform was created collaboratively by Emily Öhman and Kaisla Kajava, with Kaisla Kajava providing the final technical solution. The first draft of the manuscript was written by Emily Öhman with subsequent revisions a joint effort.

[5] http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

[6] https://creativecommons.org/licenses/by/4.0/

III Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. Creating a Dataset for Multilingual Fine-grained Emotion-detection Using Gamification-based Annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, 2018.

This update to the previous paper was conceptualized and created by Emily Öhman with input from Jörg Tiedemann and Timo Honkela. Parts of the evaluation were based on work conducted for Kaisla Kajava's MA thesis work supervised by Emily Öhman. The first draft was written by Emily Öhman and commented on by Jörg Tiedemann and Timo Honkela.

IV Emily Öhman. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, 2020.

This paper discusses annotation quality in relation to annotator motivation. This paper was for all parts completed solely by Emily Öhman.

V Emily Öhman, Timo Honkela, and Jörg Tiedemann. The Challenges of Multi-dimensional Sentiment Analysis Across Languages. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, Osaka, Japan, 2016.

Material preparation, data collection and analysis were performed by Emily Öhman, with input from Jörg Tiedemann and Timo Honkela on concept, methodology, and source data. The first draft of the manuscript was written by Emily Öhman and later revised in collaboration with Jörg Tiedemann and Timo Honkela.

VI Kaisla, Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann. Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, 2020.

This paper was based on Kaisla Kajava's graduate thesis work, supervised by Emily Öhman. Kajava is therefore listed as first author. The paper was written by Emily Öhman. Hui Piao provided input on the translation theory parts of

the paper and Jörg Tiedemann helped with the final revision of the paper.

VII Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference of Computational Linguistics (COLING 2020)*, Barcelona, Spain, 2020.
This paper was the idea of Emily Öhman, with input from Jörg Tiedemann. Marc Pàmies and Kaisla Kajava assisted with the fine-tuning of BERT and dataset evaluations. Kaisla Kajava also provided augmentations to some of the datasets from data collected for her MA thesis. The analysis was conducted by Emily Öhman. The paper was written by Emily Öhman with revisions and input from the other authors.

VIII Marc Pàmies, Emily Öhman, Kaisla Kajava, and Jörg Tiedemann. LT@Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT? *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain, 2020.
The idea for this paper came from Jörg Tiedemann as a starting point for Marc Pàmies' MA thesis work, which was supervised by Emily Öhman. Except for the sentiment analysis part, which was the work of Emily Öhman alone, the final technical solution was fine-tuned by Marc Pàmies. Parts of the analysis and evaluations were conducted by Kaisla Kajava and Emily Öhman. All authors contributed with ideas, analysis, and code. The first draft was mostly written by Marc Pàmies with input from Emily Öhman, and the final version was revised, read, and approved by all.

IX Juha Koljonen, Emily Öhman, Mikko Mattila, and Pertti Ahonen. Strength and intensity of sentiments and emotions in party manifestos: Finland, 1945 to 2019. Submitting to *Scandinavian Political Studies*.
The concept for this paper came from Juha Koljonen and Pertti Ahonen. Emily Öhman served as an expert on sentiment analysis and produced the emotion and sentiment analysis lexicons, code, and results, and wrote the sentiment analysis and method sections of the paper. The other parts of the paper were written by Juha

Koljonen with significant input from Pertti Ahonen and Mikko Mattila. Mikko Mattila also conducted the regression analyses.

X Emily Öhman and Riikka Rossi. Affect and Emotions in Finnish Literature: Combining Qualitative and Quantitative Approaches Submitted to *Proceedings of the European Association for Digital Humanities Conference (EADH 2021)*
The idea and original draft was the work of Emily Öhman. The final version contains significant input from Riikka Rossi on affect in literature. All quantitative work was conducted by Emily Öhman, and most qualitative work by Riikka Rossi.

Chapter 2

Emotions: Theory and Language

Sentiment analysis is the automated extraction and study of sentiments, emotions, or affects, in different modalities. Although emotions and sentiments are expressed via facial expressions and posture as much as words and gestures (Alm, 2012; Reilly and Seibert, 2003), in practice the modality is often text as is the case with my own research as well. Sentiment analysis differs from traditional studies into affect in, for example, literature, in its quantitative rather than qualitative approach. Sentiment analysis is typically seen as a part of computational linguistics and language technology, and sometimes computer science. In some cases it also falls under digital humanities and other transdisciplinary fields, especially when it is applied to real-world texts from traditional humanities fields such as history, literature, and ethnology.

Traditionally, sentiment analysis was just that: **sentiment** analysis. However, since two sentences can contain nearly identical information and thus both be assigned the same sentiment category (i.e. positive or negative), but in reality express very different emotions, there is a need for more fine-grained categorization in many applications. To illustrate this problem, Staiano and Guerini (2014) use the *negative* examples "I'm so miserable, I dropped my iPhone in the water and now it's not working anymore" for *sadness*, and "I am very upset, my new iPhone keeps not working!" for *anger*. Even though both demonstrate negative emotions, for *Apple* only the latter is useful or relevant for "buzz monitoring" (Staiano and Guerini, 2014).

The one thing all emotion analysis studies, both quantitative and qualitative, have in common, is that they deal with human feelings. These feelings can be defined differently, using different psychological, or even physiological, theories of emotion and labeled as affect, feeling, emotion, sentiment, or opinion. What exactly these terms mean is interpreted differently in different fields and sometimes even between researchers in the same field. Therefore, in this chapter I try to explain the different terms that relate to feelings, the focal point of sentiment analysis, and give a brief overview of different theories of emotion, how language relates to emotions and how well they are preserved in translation.

2.1 Defining the Terms

Emotions have been historically termed *passions*, *appetites*, *desires*, *feelings*, *sentiments*, *moods*, *humours*, *temperaments*, *attitudes*, *confused perceptions*, *distorted judgments*, *urges*, and many others (Maia and Santos, 2018). Sometimes these terms have been considered synonymous with *emotion*, at other times the terms have been deliberately chosen for ideological reasons, or to specify or generalize what is meant by the terms. Even in current scientific literature, emotion-related terms are often used interchangeably with little concern for congruence with previous research or even dictionary definitions (see e.g. Munezero et al. 2014).

The different terms used to describe different types of feelings can be confusing. In natural language processing (NLP), sentiment analysis often refers to simple binary or ternary polarities, i.e. *positive*, *negative*, and *neutral*, commonly on a sliding scale from -1.0 (negative) to 1.0 (positive), with *neutral* at 0. For emotion detection, more fine-grained emotions are used, such as *happiness*, *sadness*, *fear*, *anger*, etc. Most native speakers understand *sentiment* to refer to a longer lasting, more general, emotion, i.e. *love* compared to *happiness*. Affect, on the other hand, is prepersonal as contrasted with feelings that are personal and emotions that are social (Shouse, 2005).

Figure 2.1: Defining emotion-related terms on a continuum from most subconscious to most conscious.

The “Handbook of Emotions” (Barrett et al., 2016) – perhaps the most authoritative source on emotions in psychology – states that “[t]heorists of emotions disagree vigorously on what emotions are” (Scarantino, 2016, 5). The handbook presents three main traditions of emotion: *the feeling tradition*, *the motivational tradition* and *the evaluative tradition* which are further subdivided into at least 15 separate approaches and -isms.

In the *feeling tradition* emotions are considered to be feelings of distinctive types and

feelings conscious experiences (Scarantino, 2016). Many famous philosophers fall under this tradition. Among them Hume, Descartes, Russell, James, and Lange (Scarantino, 2016). The feeling tradition, with emotions equaling feelings, has largely been marginalized in the 20th century (Scarantino, 2016).

For those who subscribe to the *motivational tradition* emotions are considered motivational states or patterns of behavior of a distinctive type (Scarantino, 2016). Ekman and Plutchik, who will be discussed in more detail later in this section, are considered to be a part of this tradition in its more contemporary form of *basic emotion theory and social constructivism* following the demise of behaviorism (Scarantino, 2016). Although the motivational tradition is still highly influential in the "contemporary emotion debate" (Scarantino, 2016, 23) it shares the challenge of differentiation with the feeling tradition.

The *evaluative tradition* considers emotions to be distinguished from each other by the involved evaluations, i.e. an interpretation of circumstances (Scarantino, 2016). Scherer, who is also mentioned later in this section, is a follower of the evaluative tradition, specifically the *appraisal theory* branch.

Scarantino (2016, 37) states that "the emotion community continues to be divided on the nature of emotions, [and] on the terminology suitable to describe them". With all these experts in disagreement, and the field of emotion research still in flux, I am also utilizing everyday dictionaries (mainly Merriam-Webster) to tease out the meaning of these different terms as much as is possible. Although the psychologists' definitions on the topic of emotion terms are perhaps the more authoritative ones, the dictionary definition too is relevant for a few reasons: (1) the vastly different ways in which these terms are used in different fields of science means that studying just one or two psychologists on the topic is not likely to give a definitive answer (if it is even possible for one to exist), (2) in many fields "sentiment analysis" is an umbrella term used for emotion detection and studies of affect as well as what is understood as sentiment analysis in language technology, and (3) for many laymen, the authoritative definition would indeed be the dictionary definition, something which is likely to influence those in particular who study sentiment analysis from a point of view unrelated to psychology.

Indeed, *feelings* are defined by Merriam-Webster as an *emotional state or reaction* (Merriam-Webster Online, 2009), a definition comparable to Shouse (2005) who adds

a sensation that has been checked against previous experience and labeled whereas *emotions* are the projection or display of a *feeling*. *Emotions* are social in the sense that they can be feigned and they change when our social circumstances change.

Sentiment is an attitude, thought, or judgment prompted by feeling or an idea colored by emotion (Merriam-Webster Online, 2009) Here, Merriam-Webster again seems to subscribe to the *feeling tradition*. Sentiments are conscious and fleeting emotional dispositions that have developed over time. *Sentiments* are sometimes used interchangeably with *opinions* as they are very closely linked (Kim and Hovy, 2004), also demonstrated by how *opinion mining* is sometimes used synonymously with sentiment analysis. The main difference between them is that *opinions* are generally personal interpretations of emotions, feelings, or sentiments (Munezero et al., 2014).

Affect has a few differing definitions with Merriam-Webster defining it as the conscious subjective aspect of an emotion considered apart from bodily changes, but Shouse (2005) defines it as a non-conscious experience of intensity and the most abstract of the terms (see figure 2.1). If going by Shouse's definition, which in turn is based on Damasio (1994), *affect* is the most basic and subconscious of these terms. If going by the Merriam-Webster definition *affect* should be considered a subjectively experienced emotion and therefore between emotion and sentiment in terms of consciousness.

In *affective computing* (Picard, 1997) in particular, *affect* is often studied in conjunction with *valence* and *arousal*. *Arousal* is generally considered to be synonymous with *intensity* and *valence* is the extent to which an emotion is positive or negative. These concepts are inter-related. In affective computing, affect is used to refer to all the related and connected concepts that have been mentioned in this section, including emotion, mood, feelings, personality, attitude, polarity etc. (Picard, 1997; Alm, 2012). This is likely part of why all these different terms are used interchangeably in some studies and mean somewhat different things in different fields.

The fact that there is overlap in the use of these different terms does not mean that the different views are pragmatically incompatible: "Pragmatically, different views of affect complement each other and jointly create a basis for understanding affective language phenomena" (Alm, 2010). But, because some fields, especially those where the focus is on downstream applications of sentiment analysis and emotion detection, do not

distinguish between sentiments and emotions (see e.g. Widmann 2019 for an example) it is imperative to clarify what is meant by these terms in this dissertation. The terms that are used are *emotion* and *sentiment*. In my work I take *sentiment* to mean general *positive*, *negative*, and *neutral* categories and as more long-lasting than emotions. *Emotion* stands for a more fine-grained division into specific sentiment-provoking feelings such as *anger*, *fear*, *joy*, *sadness*, *disgust*, etc. that are more physiological and thus "basic" in nature.

2.2 Theories of Emotion

Philosophers and scientists have long debated the nature of emotions and the classification of these emotions. One of the first written works on the topic date back to the ancient Greeks and Romans, Chrysippus and Cicero in particular, and Aristotle divided feelings into *passions* and *actions* (Barrett et al., 2016). However, much of the foundation for the modern study of human emotions was laid by Darwin in his book *The Expression of Emotions in Man and Animals* (Darwin, 1872).

For decades experts debated whether emotions are innate and biological or whether emotions were a cultural experience and subject to cultural influence (see e.g. Tomkins 1962; Scheper-Hughes 1984). In 1971 Ekman published his findings on universal emotions (Ekman and Friesen, 1971; Ekman, 1971). He suggested that there were six basic emotions shared by all humans: *fear, anger, disgust, sadness, happiness, and surprise*. Although some additions to this core group of six emotions have been made later, including by Ekman himself, these additions have not become as universally accepted as the original, partly due to the fact that the later research was not pan-cultural (Ekman, 1992), which was crucial to Ekman's theories being accepted in the first place.

Much of modern research on emotions is at least to some extent based on the work of Ekman. This includes the work of Robert Plutchik, whose Wheel of Emotions is the theory of emotions that this dissertation uses as a basis for classifying emotions. Both scholars' backgrounds are in psychology, which is reflected in their theories, and indeed in nearly all theories of emotion. It should, however, be noted that neither theory is optimized for the analysis of text.

Most recently, Keltner and Cowen (Keltner et al., 2019b,a; Cowen and Keltner, 2017; Cowen et al., 2019) have tried to tackle the categorization of emotions in a number of studies and have come up with an emotion categorization consisting of 27 distinct emotions by studying emotion responses to a number of different stimuli such as videos, music, facial expressions, speech prosody and even nonverbal vocalization (Cowen et al., 2019; Demszky et al., 2020). This is the emotion annotation scheme partly relied on in GoEmotions (Demszky et al., 2020). But although this categorization has its benefits, it too suffers from some of the same issues other categorization schemes suffer from, namely that it is not designed for emotion detection in **text** specifically.

2.2.1 Plutchik's Wheel of Emotions

Plutchik's wheel (Plutchik, 1980) is an augmentation of Ekman's research. The core emotions are the same with the addition of *trust*[1] and *anticipation*, where *happiness* has been replaced by *joy*[2]. Plutchik expands on Ekman's theories and presents emotions on a three-dimensional model based on basic-to-complex[3] categories and dimensionality. In his wheel, emotions are arranged in concentric circles in a flower-like shape with the petals representing emotion categories that can be combined in various ways (see Figure 2.2). The wheel has 8 core emotions, and with the combinations between dyads result in a total of 56 emotions at the same intensity level. The wheel also maps three intensity levels to all the core emotions. Plutchik (1984) also describes his wheel as a cone, which clarifies how emotions become less distinguishable at lower intensities (See 2.3).

It should be noted that Plutchik himself views the valence of *anger* as positive and *surprise* as negative. In a majority of sentiment analysis and emotion detection systems, *anger* is considered negative and *surprise* neutral or it is not included (see e.g Mohammad et al.). From a practical point of view, it makes sense to consider *anger* negative

[1] In some of Plutchik's work, e.g. Plutchik (1984), *anticipation* is referred to as *acceptance*.

[2] In some of Plutchik's work, e.g. Plutchik (1984), *disgust* is referred to as *rejection* (Plutchik, 1984)

[3] See Plutchik (1980); Frijda (1988) and Ekman (1992) for a discussion on basic emotions in comparison to complex.

[4] Source of original figure: `https://en.wikipedia.org/wiki/Contrasting_and_categorization_of_emotions`.

[5] Figure from Juslin (2013), CC BY 3.0.

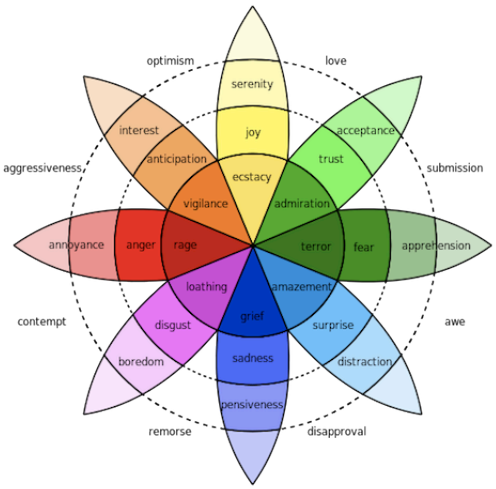

Figure 2.2: Plutchik's wheel of emotions.[4]

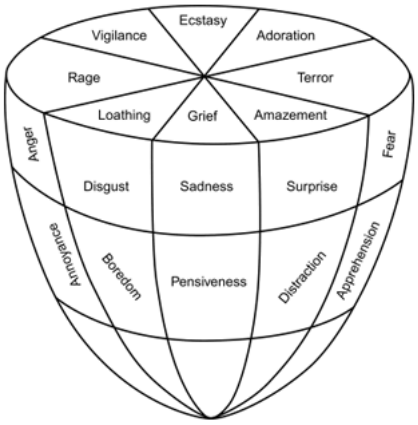

Figure 2.3: Plutchik's multidimensional cone model.[5]

because if a customer review comes across as angry, it is unlikely that the review is positive. For *surprise*, the problem seems to be that it tends to be the hardest category to classify correctly (Zampieri et al., 2019a; Mohammad et al.; Kirkland and Cunningham, 2012). *Surprise* is quite easy to recognize from facial expressions (from where its inclusion in the core emotions stems from), but it is a difficult emotion to express in text.

Plutchik's work has influenced several new theories of emotion. Perhaps chief among them for sentiment and emotion analysis purposes is *The Hourglass of Emotions (Sen-*

ticNet) (Cambria et al., 2012; Cambria and Hussain, 2012; Cambria et al., 2013). In this theory, Plutchik's multidimensional cone model and wheel are reworked to show the change in emotional intensity on a Gaussian curve to better fit with human-computer interaction and affect studies (see figure 2.4). In it emotions are categorized into four *sentic* dimensions where joy, calmness, pleasantness and eagerness are considered positive emotions, and sadness, anger, disgust and fear as negative.

Figure 2.4: The Hourglass of Emotions[6]

Although there have been recent attempts to develop new emotion classification schemes (such as the already mentioned Cambria et al. (2011), but also Cowen and Keltner (2017), Cowen et al. (2019), Keltner et al. (2019b), Keltner et al. (2019a), and Demszky et al. (2020)) where emotion categorization is explored in different modalities,

[6]Image from Wikipedia `https://en.wikipedia.org/wiki/Emotion_classification`, CC BY-SA 4.0.

and in the case of Demszky et al. (2020) also applied in emotion detection, Plutchik's eight core emotions remain the most used in language technology (see e.g. Bostan and Klinger 2018).

The theory that most of the research presented in this dissertation is based on is an adaptation of Plutchik's wheel and his core emotions, and therefore going into great detail about other models and theories is beyond the scope of this dissertation. Plutchik's theory was chosen both for practical and theoretical reasons. It is the most commonly used theory in fine-grained sentiment analysis and is the theoretical basis for the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013). The different intensity levels allow intensity regression (see also Akhtar et al. 2020). Furthermore, although the theory was not developed for the medium of text, it has successfully been used in thousands of papers for text-based sentiment analysis. Hence, for paper V, chronologically my first paper, I chose to use Plutchik because it was so widely used. Subsequently I wanted to be able to compare results from that study with results from my later papers, as well as with the findings and results of other scholars.

2.3 Language and Emotion

It has been theorized that language and emotions co-evolved. Jablonka et al. (2012) suggests that human-specific social emotions are linked to information-sharing and the evolution of language enhanced the inhibitory control of emotions leading to a mentality that differs from that of other apes. The relevance of emotions for language can be considered from three perspectives (Foolen, 2012, 349 - abstract):

1. The conceptualization of emotions
2. The expression of emotions
3. The grounding of language

The conceptualization of emotions refers to the emotional lexicon, i.e. words that express emotions. These are usually nouns, adjectives, and verbs, but can also be prepositions (e.g. in how certain prepositions are used significantly more with emotion words).

Emotions are expressed on all linguistic levels including phonological, morphological, lexical, syntactic etc. Grounding refers to emotions being a precondition to the functioning of language (Foolen, 2012).

Since emotions are conceptualized using words, the most immediate concern for sentiment analysis tasks is at the lexical level. What level of granularity, i.e. word-, sentence-, or document-level, is relevant and useful depends on the approach and system design. With more and more advanced classification methods for machine learning available an increased selection of linguistic levels can be included, e.g. syntactic, morphological, and other aspects such as metaphor, metonymy, irony, and sarcasm.

Emotions are a topic that link many different fields. Emotions are studied from vastly different perspectives including linguistic, psychological, ethnological, anthropological, philosophical, medical and neurological, as well as computational. Although beyond the scope of this thesis, it should be noted that emotions have been studied quite thoroughly in psychology, but there are still many competing and interlinked perspectives on what emotions really are (Cornelius, 2000; Alm, 2012).

2.4 Emotions in Translation

Research shows that affect categories are quite universal (Cowen et al., 2019; Scherer and Wallbott, 1994), and therefore emotion categories should be relatively stable even when a text is translated. However, in practice there are linguistic and cultural limitations to how well emotions translate, both in terms of how easy they are to translate in terms of words and constructs, and culturally in how emotions are presented with words. Although issues specific to relatedness exist when translating closely related languages (such as false friends and overly literal translations (Norberg and Stachl-Peier, 2014; Gillberg, 2005)), it is probably safe to assume that the translation of emotion words works best between genealogically related languages spoken in culturally similar regions (Hægermark, 1997).

Wierzbicka (1999) believes that even the term *emotion* is Anglo-centric. However, her research into Natural Semantic Metalanguage suggests that despite the infinite set of possibilities for emotions to be lexicalized in different societies, they can also be

translated and understood by all humans (Wierzbicka, 1999; Santos and Maia, 2018). Such a list of *semantic primitives* is still only an analytical tool since there are no one-to-one translations of emotion words between different languages (Santos and Maia, 2018; Maia and Santos, 2018).

Some languages and cultures shy away from expressing emotions by words, and place a bigger emphasis on actions instead. In many vastly different cultures, emotional outbursts are considered *tacky*, whereas in others the lack of such outbursts can give an impressions of coldness. These, and other, issues make translating emotions difficult. Using the same sentiment analysis model on two different languages, even if the text is the same in content, can yield very different results.

Sentiments can be very culturally dependent and even texts that have been translated correctly can be marked as having a different sentiment than what the original text intended when evaluated by native speakers. Mohammad et al. (2016) as well as Salameh et al. (2015) list mistranslation (and the related "translated to a related word"), cultural differences, and different sense distributions between languages as the primary reason for errors when automatically translating entries in a sentiment lexicon. In paper I, I translated the NRC Emotion Lexicon (Mohammad and Turney, 2013) and found similar issues in the Finnish translations of English sentiment and emotion words.

Greek has no simple expression for *frustration* (Pavlenko, 2008). On the other hand the Dholuo language has the expression *maof*, which is described as "the feeling of desiring to see relatives and friends that have not been seen for too long and is by extension transferred to other things" (Omondi, 1997; Satthar, 2018). Some languages shy away from saying e.g. *I love you*, despite the words to express that sentiment existing in the language. Therefore, if an original text in such a language does express that sentiment explicitly, it seems that the literal English translation might not sufficiently express the intensity of the sentiment. Similarly, if translating from a language like English into such a love-shying language, it stands to reason that explicitly expressing that sentiment, would be a cultural incongruity.

Although Ekman in his studies on emotion recognition claimed to have found universal emotions[7], Shimoda et al. (1978) found that when English, Italian, and Japanese

[7]Ekman does discuss *Japanese display rules*.

subjects were tasked with recognizing facial expressions of emotions, there were significant differences in how the different nationalities fared. The Italian subjects failed to recognize Japanese expressions above chance accuracy (Shimoda et al., 1978, 177). Imada (1989) found that there are many differences between how English and Japanese express *anxiety*, *fear*, and *depression* in text.

In previous computational work, sentiment and emotion preservation in translation has been studied mainly as a machine translation issue. For example, Mohammad et al. (2016) tested how translation alters sentiment by translating a low-resource language text into English and applying English state-of-the-art sentiment analysis systems on the translated data. They found that automatic sentiment analysis of automatic translations outperformed the manual sentiment annotations of the automatically translated text. A common issue was the loss of sentiment when translated, i.e. a previously positive or negative expression became neutral after translation.

Lohar et al. (2017) assumed that sentiments are preserved poorly when translation quality is low and with this in mind trained a classifier to investigate how these two issues are intertwined. Their main finding was that splitting the machine translation training data into sentiment classes improved sentiment preservation in the target language.

A link between machine translation and annotation quality has been discussed by at least Salameh et al. (2015) and Mohammad et al. (2016). Salameh et al. (2015) discovered, somewhat surprisingly, that the accuracy of sentiment classification of machine-translated texts was higher than that of manual sentiment annotation of the same text, despite losing sentiment information in translation and increasing the number of items predicted as *neutral*. They speculated that this was due to how machine translation errors impacted the annotators' perception of the automatically translated text.

It seems clear that the emotion lexicons in different languages are far from identical. Some languages do not have words for certain emotions or do not differentiate between emotions that are separate categories in other languages (Foolen, 2012). The differences in the surface realizations in different languages do not necessarily mean that emotions are not present in a similar way and intensity in translations, but it makes exploring the preservation of emotions in translation quantitatively more difficult. This is something that any study dealing with emotions as expressed in language and text needs to consider.

Chapter 3

Datasets and Annotation

Annotations are a crucial part of both data-driven methods and lexicon-based methods and their evaluation (these methods will be discussed in chapter 4). For lexicon-based methods, the target of the annotations are usually words in order to create lexicons. For annotated datasets, which are required for supervised machine learning, the target of the annotations vary with the domain of the data. The granularity of annotations can also vary: e.g. phrase, sentence, paragraph, or document level. In practice it is nowadays often social-media post level, and therefore the length varies between different social media platforms. These annotations can be acquired manually using human annotators, or automatically using a feature of the data such as hashtags or star ratings.

The manual annotation process requires annotators, an annotation platform, as well as data or dictionaries to be annotated. In this chapter I will first talk about annotation and emotion annotation in particular, as well as annotation projection before presenting existing datasets starting with sentiment-only datasets, and then both emotion lexicons and emotions datasets of which I will discuss a few in more detail. I also present existing annotation tools.

3.1 The Annotation of Emotions

Supervised machine learning tasks (see section 4.3) rely on annotated data, but the annotation of datasets can be very costly and time consuming (Andreevskaia and Bergler, 2007; Devitt and Ahmad, 2008). Although there are ways to annotate data automatically, e.g. by relying on */s* to label social media posts as *sarcasm*[1] and use these as training data (Khodak et al., 2017), the annotation task is commonly performed by human annotators despite the exponential increase in cost, as humans are much better at understanding natural language than computers.

Crowd-sourcing can often be a cheaper alternative to hiring expert annotators, and has been used successfully in several projects to create different types of annotated datasets, including sentiment and emotion annotated ones (Turney, 2002; Greenhill et al., 2014; Mohammad and Turney, 2010, 2013; Öhman, 2020). One issue with using non-

[1] In most social media, but *Reddit* in particular, /s is commonly appended after a sarcastic post to denote sarcasm where there is a risk of the message of the post otherwise being taken seriously.

experts to solicit annotations is that there is a risk of the quality suffering. This risk can be mitigated by carefully controlling selection criteria (Hsueh et al., 2009), and being aware of typically difficult instances to annotate such as requests, the speaker's emotional state, neutral reporting of positive or negative facts and so forth (Mohammad, 2016).

Due to the subjectivity of most annotation tasks, however, humans do not always agree with each other on how to annotate. Whether a tweet, for example, expresses surprise or fear can be ambiguous, and heavily depends on the reader's interpretation. There is rarely one single reading that can be judged as being correct (Campbell, 2004) as the annotation task is generally highly subjective even with careful annotator guidelines (Mohammad, 2016). It should also be noted that emotions do not have distinct and clear boundaries that separate them and they often occur together with other emotions (Mohammad and Turney, 2013).

Previous annotation tasks have shown that even with binary or ternary classification schemes, human annotators agree only about 70-80% of the time and the more categories there are, the harder it becomes for annotators to agree (Bayerl and Paul, 2011; Boland et al., 2013; Mozetič et al., 2016). For example, when creating the DENS dataset (Liu et al., 2019), only 21% of their annotations had consensus between all annotators with 73.5% having to resort to majority agreement, and a further 5.5% could not be agreed upon and were left to expert annotators to be resolved.

Aspects that have been shown to influence inter-annotator agreement are: the domain of the data being annotated, the number of labels and categories in the annotation scheme, the training and guidelines of the annotators as well as the intensity of that training, how many annotators there are in total, for what purpose the annotations are, and of course the method used to calculate the inter-annotator agreement (Bayerl and Paul, 2011). Some of these considerations are more about the mathematical aspect of calculating the agreement, including the point about number of annotators. Interestingly, Bayerl and Paul (2011) found that increasing training improved agreement scores, when Mohammad (2016); Mohammad and Turney (2013) found that minimal guidelines improved agreement scores because over-training annotators led to confusion and apprehension in judgement tasks. Which approach is better for improving agreement scores,

is probably related to the annotation task and annotator expertise. Additionally, Bayerl and Paul (2011) suggests that when resolving conflicting annotations, it is better to use the majority annotation rather than attempt to find a consensus solution.

One of the reasons that make emotion annotation difficult is the modality, i.e. text versus speech versus video and so forth. In typical human interactions, emotions are expressed through multiple modalities simultaneously Cowen et al. (2019), but in annotation tasks, the focus is usually on one single modality, most commonly text. Emotions are expressed in text through various means that are limited by linguistic, cultural, and social constraints. In contrast, the most commonly used emotion annotation schemes are based on psychological theories that are in turn based on human interactions, not text. Therefore annotating outside the originally intended modality or environment makes the emotion annotation task harder.

Some emotions are also harder to detect and recognize. Demszky et al. (2020) show that the emotions of *admiration, approval, annoyance* and *gratitude* had the highest interrater correlations at around 0.6, and *grief, relief, pride, nervousness, embarrassment* had the lowest interrater correlations between 0-0.2, with a vast majority of emotions falling in the range of 0.3-0.5 for interrater correlation. Liu et al. (2019) too notes that they had difficulties with the categories of *disgust* and *surprise*. In their case, *disgust* was such a small category that it should be discarded, and surprise was such a noisy category, that it too should be discarded. Alm (2010), who had similar observations regarding *surprise*[2] as Liu et al. (2019), speculates that this is because *surprise* is characterized in text by direct speech and unexpected observations. On the other hand, she found that *fear*[3] was often marked by words directly associated with *fear*. Sentences that contained outright affect words were more likely to have high inter-annotator agreement (Alm, 2010).

Noisy annotations, i.e. annotations that differ from the majority within a dataset, are a problem for machine learning approaches, since they are feeding the machine conflicting information which means that the model will exhibit what in a human would be called confusion. Noise can also more generally refer to irrelevant information or ran-

[2] *Surprised* in her annotation scheme.
[3] *Fearful* in her annotation scheme.

domness in a dataset. Noise in many cases can also refer to different readings and the inherent ambiguity of the data. Therefore, in order not to overfit a model (see chapter 4) by tying it to just one reading, multilabel classification can help by allowing these subjective readings to be a part of the training process.

Annotation quality is usually measured using different methods to calculate inter-annotator agreement, sometimes referred to as interrater correlation. The most commonly used method is Cohen's Kappa. Typical inter-annotator agreement varies and is affected by the type of annotation, experience level of annotators, and the granularity of the annotations (Bermingham and Smeaton, 2009; Ng et al., 1999). In a straightforward image annotation task annotation agreements were around 90% (Nowak and Rüger, 2010), however, for a sentence-level sentiment annotation task, inter-annotator agreements stayed at around 57% with Krippendorf's α scores of α = 0.4219, well below even tentative reliability (α = 0.666...) (Bermingham and Smeaton, 2009). In a word sense disambiguation task, κ values were around 0.3, indicating very low agreement (Ng et al., 1999).

If human annotators find it this hard to agree, it seems unreasonable to expect computers to perform much better. Especially since if computers are trained on human annotated data, these disagreements can easily confuse a computer in the learning process reducing the accuracy of predictions further. Although computers are getting better and better at natural language understanding and machine learning is progressing fast, human annotators are unlikely to ever achieve better agreement rates. Annotator training and removal of noise from data can improve the quality of the dataset (Mohammad, 2016), but Bermingham and Smeaton (2009) suggest that especially with sentiment annotation tasks, sentence-level annotations might just be too fine-grained since they found that agreement increased by simulating document-level judgements. They also suggest that instead of the fairly commonly included *mixed* category for sentiments, to rather have an *indeterminate* cateogory as many annotators reported feeling uneasy over their judgements.

Often datasets consist of data from the same source, which tends to render them useful in that same domain only. This domain specificity means that, in general, every classification task for a different domain needs its own annotated dataset further increas-

ing the cost of supervised machine learning. If you are training a model to detect cancer from x-rays, it makes sense to use x-rays as training data. This is true for sentiment analysis as well; if you want to detect emotions in tweets, a model trained on Amazon product reviews or IMDB movie reviews is not going to be as helpful (although possibly not completely useless) as a model trained directly on other tweets.

A related issue is that of context-dependence. Especially when dealing with fine-grained data such as sentences, it is much easier for a human to annotate when they can see the sentence in context. Why then not always annotate everything within context? Unfortunately, this renders the annotation heavily context-dependent. If you think of a situation where the annotator only sees the text *Come on! Let's go!* it is very difficult to say definitively what the emotion expressed is without context. In this case it could be *fear, anticipation, anger, sadness...*

On the other hand we can have many different situations where this could be expressed and if the annotator has context, then the annotation is only valid for that particular context. To use the previous example, if the context is (a) seeing something scary, or (b) seeing that there is no one in line at your favorite ride at Disneyland, it completely alters the associated emotions. What if the data point is there twice, once for both contexts? Then we would have two lines that are identical without context, annotated with opposing and, generally, incompatible emotion categories because of context, confusing the learning process of a machine learning model.

Context can also lead to the effect of double weighting fine-grained annotations (Boland et al., 2013). That is, if a sentence is annotated in context and is only negative in that context, when the contextual sentences are then also annotated as negative, the first sentence unduly increases the overall negativity of those sentences. If sentences are classified in context, it makes sense to annotate them in context, but if not, whether to have context-sensitive annotations or not is something that should be carefully considered separately for each project.

Annotating for intensity or valence has its own additional difficulties. It is very difficult, if not nearly impossible, for humans to give value scores for annotations. Despite this, sometimes simple rating scales are used, where people are asked to rate items on a scale. One example of such as scale is the well-known Likert scale of agreement (Lik-

ert, 1932). Not only will two people have different ideas on how those scores relate to a word and how the words relate to each other, but self-consistency is also difficult. There are also issues with "scale region bias", which simply means that certain regions of the scale for scoring will be over-represented.

Paired comparisons are sometimes used where annotators compare two items to determine which has more of the given property. This type of comparison works very well and gives accurate scores for items or words, but it requires an exponential number of annotations (n^2, where n is the number of items to be annotated according to Kiritchenko and Mohammad (2017)).

An improved version of paired comparisons, best-worst-scaling (Louviere and Woodworth, 1991) works much in the same way except instead of two items to be compared, multiple items are simultaneously ranked in relation to each other. Typically four items at a time are compared. The required number of annotation instances is then only $1.5-2n$ compared to n^2 with paired comparisons.

Kiritchenko and Mohammad (2016, 2017) showed that the reliability obtained by using 10 annotations per item with rating scales ($10n$) can be matched with only three total annotations with best-worst-scaling ($3n$). Furthermore the results are less biased, and contain more discriminating annotations while preserving the comparative nature of them (Kiritchenko and Mohammad, 2017).

Strapparava and Mihalcea (2010, 30) outright state that "emotion annotation is difficult". The same conclusion is reached by many other researchers who have evaluated annotations (Mohammad, 2016; Wiebe and Riloff, 2005; Andreevskaia and Bergler, 2007; Munezero et al., 2014; Bermingham and Smeaton, 2009).

3.2 Annotation Projection

Annotation projection is the process of taking the annotations of one dataset and projecting them onto an aligned parallel dataset. Usually, but not always, the annotated dataset is in one language and the dataset to which the annotations are projected are in a different language. They can also be, for example, a monolingual paraphrase corpus or a monolingual simplification corpus. In practice these datasets are almost exclusively

some type of parallel corpora for which annotations have been solicited for English only.

For example, Tiedemann (2014) showed that the limited success of annotation projection for parser induction across languages in previous work was mainly due to evaluation with incompatible annotation schemes, and that these obstacles could be overcome with harmonized annotations across the languages. Yarowsky et al. (2001) showed that for part-of-speech tagging, English tools and resources can successfully be harnessed for low-resource languages with the help of annotation projection.

For annotation projection for sentiment analysis Banea et al. (2011) found that right after having a manually annotated dataset in the target language, the second best option was to project annotations from a resource-rich language to a low-resource language. This of course requires that some kind of connection can be established between the two languages, either by using parallel corpora or machine or manual translation. One of the earliest systems to do this in practice was Mihalcea et al. (2007) with subjectivity labels. They achieved f-scores between 0.4366 and 0.4793 with a rule-based classifier and 0.6785 with a machine learning approach for Romanian data with subjectivity annotations translated from English (these evaluation methods will be discussed in more detail in section 4.1.1).

More recently, Balahur and Turchi (2014) also used machine translation to transfer sentiment information from English to lower-resource languages (Spanish, French, and German). These projections were then evaluated using SVMs (see section 4.3) with Sequential Minimal Optimization. They came to the conclusion that machine translation systems are robust enough to be able to aid in annotation projection and the creation of tools and resources for low-resource languages, but that this is highly dependent on translation quality.

On the other hand, Barnes et al. (2018) compared annotation projection and machine translation in the creation of datasets for low-resource languages. Their dataset was annotated into for classes: *strong negative*, *weak negative*, *weak positive* and *strong positive*. Barnes et al. (2018) found that machine translation performed much better than the projected annotations, both with binary classification and 4-class classification. For machine learning their model produced macro f1 scores of 0.79 for binary and 0.52 for 4-class, and projection yielded scores of 0.69 for binary, and 0.44 for 4-class. Although

machine translation worked better in this case, annotation projection still yielded results that indicate that a projected dataset is useful if automatic translation can not be done[4], but parallel data is required in that case. As the quality of machine translation varies greatly between language pairs, and tend to be worse for low-resource languages due to lack of suitable training data, annotation projection is likely still the better alternative in most cases.

To summarize, it seems that annotation projection can be a viable option in many cases, but that when possible it should be compared to annotation projection of labels to a machine translated version of the original dataset. To my knowledge though, no previous study exists which compares multilabel, multiclass manual annotations with annotation projection and machine translation.

3.3 Sentiment and Emotion Lexicons and Resources

Sentiment and emotion lexicons can list words and whether or not those words are associated with an emotion or sentiment (or multiple). Dictionaries and thesauruses can be used to create sentiment lexicons (Hovy, 2010; Kim and Hovy, 2004) as can human annotators (Mohammad and Turney, 2013), and using corpora by calculating co-occurence with annotated seed words linked with emotion (Esuli and Sebastiani, 2006; Turney, 2002; Satthar, 2018; Kaji and Kitsuregawa, 2007).

In addition to listing the sentiment, a lexicon can also list the intensity of said sentiment (Mohammad and Bravo-Marquez, 2017; Kiritchenko and Mohammad, 2016). Kiritchenko and Mohammad (2016) use best-worst scaling (BWS) to create an intensity emotion lexicon. BWS allows annotators to rank words according to intensity and is much more reliable than other common approaches such as Rating Scales (RS) (Kiritchenko and Mohammad, 2017) as in effect the intensities do not rely on subjective external values of intensity, but are a function of the intensity relation between words.

Table 3.1 shows a fictional example of an emotion lexicon, where intensities have

[4]I.e. for most languages. Google Translate only covers 109 languages, and estimates suggest the number of languages in the world is between 6,500 and 8,000 (`https://translate.google.com/intl/en/about/languages/` and `ethnologue.com`).

word	emotion	intensity
promotion	joy, anticipation	0.71
furloughed	anger, fear	0.73
roller-coaster	joy, anticipation, fear	0.46

Table 3.1: Fictional example of an emotion lexicon

been obtained by BWS, that combines words of different parts-of-speech, emotion labels, and intensity scores.

Quite a few lexicons exist too, for example, the WordStat Sentiment Dictionary (Loughran and McDonald, 2011), the Sentiment Lexicon for 81 Languages (Chen and Skiena, 2014), SentiWordNet (Esuli and Sebastiani, 2006), AFINN (Nielsen, 2011), and of course the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013) and other NRC sentiment, emotion, affect, and association lexicons[5]. For an overview of **emotion lexicons** as opposed to sentiment lexicons, see table 3.3. One of the first that is still used today is the Movie Review Data (Pang et al., 2002). The MPQA Opinion Corpus (Deng and Wiebe, 2015) is also a well-known resource, particularly in academic circles. More recent datasets are the DENS (Liu et al., 2019) and GoEmotions (Demszky et al., 2020) datasets. Except for these two datasets, and the NRC Emotion Lexicon, all of the others use a sentiment annotation scheme rather than emotion annotation. That is, they only annotate for positive and negative sometimes with the addition of neutral and/or intensities. In table 3.2 I present an overview of the **sentiment** datasets, i.e. those using an annotation scheme that does not include fine-grained emotions, discussed here.

Review datasets are very common because they are easy to create. There is no need for human annotators in the usual sense because the data has already been annotated by humans when they reviewed the product. Typically, non-review datasets are likelier to use more sophisticated annotation schemes (e.g. Webb et al. (2005) and Deng and Wiebe (2015)). Except for the explicitly multilingual, all datasets are for English.

For a comprehensive analysis of six of these sentiment lexicons; see Khoo and Johnkhan (2018), however, some of their findings should be taken with a grain of salt as they compare pure binary sentiment lexicons with intensity lexicons and emotion lexi-

[5]For a full list, please visit: `http://saifmohammad.com/`.

Dataset	Size	Domain	Granularity
Stanford Sentiment Treebank	10k	Movie reviews	Review
IMDB Movie Reviews	50k	Movie reviews	Review
Paper Reviews Dataset	405	Academic papers	Paper
Sentiment140	1.6M	Tweets	Tweets
Opin-Rank Review Dataset	300k	Trip reviews	Review
Amazon Product Data	142.8M	Product reviews	Review
WordStat Sentiment Dictionary	15k	Dictionary	Words
Sentiment Lexicon for 81 Languages	varies	Dictionary	Words
MPQA	10k	News	Word, phrase
SentiWordNet	115k	WordNet	Synset
Movie Review Data	2k (main)	Movie reviews	Review
AFINN	2.5k	Tweets	Word

Table 3.2: Overview of existing resources for **sentiment** analysis

cons that have been scaled down to sentiments only. Naturally, a sentiment lexicon with sliding valence scores will give better results than a lexicon annotated as simply positive or negative, not to mention a lexicon that was first and foremost annotated for emotions, not sentiments.

I already mentioned a few emotion datasets previously. In table 3.3 I present some of the most common emotion lexicons. The NRC Emotion Lexicon used in my research and described in detail in section 3.3.1 is highlighted.

There are many differences in how these lexicons were created. For example *DepechMood* was created automatically, as was the NRC Hashtag lexicon and AffectNet. The others were human-annotated using various platforms and agreement thresholds. Strapparava et al. (2004) created WordNet Affect by cross-referencing emotions from lexicon with synsets in EmoWordNet and then expand on these by linking to related synsets. Staiano and Guerini (2014) developed DepechMood which is also aligned with EmoWordnet, and later expanded unofficially by Badaro et al. (2018) and Song et al. (2015) and officially by Araque et al. (2019).

DepechMood++ (Araque et al., 2019) also expands the originally English lexicon

Emotion Lexicon	Size	Creators
NRC Emotion Lexicon	14k	Mohammad and Turney (2013)
Fuzzy Affect	4k	Subasic and Huettner (2001)
WordNet-Affect	1k	Strapparava et al. (2004)
Affect database	2k	Neviarouskaya et al. (2010)
AffectNet	10k	Cambria et al. (2011)
NRC Hashtag	16k	Mohammad and Kiritchenko (2015)
NRC Affect	6k	Mohammad (2018b)
DepechMood	37k	Staiano and Guerini (2014)
DepechMood++	176k	Araque et al. (2019)

Table 3.3: Overview of emotion lexicons

to Italian, however, their examples show large differences for dominant emotion for the same words in different languages. For example, *criminal, criminale* has the dominant emotion *angry* in English, and *afraid* in Italian. *Rapist, stupratore* has the dominant emotion *angry* in English and *annoyed* in Italian, and *warning, allarme* has *afraid* in English and *sad* in Italian. Whether this is because of cultural differences — meaning that these words actually do have different dominant emotions due to cultural reasons — or because of weaknesses in their lexicon creation method, is unclear.

All in all, many different sentiment resources and emotion lexicons exist, the one I use in my research the most is the NRC Emotion Lexicon, described in detail in the next section. I chose this lexicon for a few reasons: (1) it was freely available and easy to access, (2) it was well documented both in terms of creation and content, (3) once I had used it for paper V, my chronologically first paper, I wanted to keep using it in order to be able to compare results from subsequent studies, and (4) most emotion datasets seemed to use the lexicon itself or the same annotation scheme in some way in the annotation process, meaning my datasets would also be comparable to those datasets.

3.3.1 The NRC Emotion Lexicon

The original NRC Emotion Lexicon (Mohammad and Turney, 2010) with 2000 English words was created from a combination of dictionary terms and phrases that were anno-

tated by humans using Mechanical Turk on Plutchik's core emotions plus positive and negative. For this lexicon, the annotators were asked to annotate for the emotion the given word **evokes**.

The augmented version of this lexicon (Mohammad and Turney, 2013) added the Ekman subset (see section 2.2 and Ekman 1992) of the WAL (WordNet Affect Lexicon) (Strapparava et al., 2004) and GI (General Inquirer) (Stone et al., 1966). This time the authors tested whether asking annotators to annotate for emotions that the terms evoked or the emotions the terms were associated with, resulted in higher inter-annotator agreement. As association produced significantly better results (except for the emotions of *fear* and *sadness*), the final lexicon[6] has 14,182 words annotated for **association** with emotions in English and translated into over 100 languages using Google Translate.

Examples of words associated with specific emotions are (Mohammad, 2018a):

- *violence* and *shout* are associated with *anger*
- *shudder* and *public speaking* are associated with *fear*
- *yummy* and *vacation* are associated with *joy*
- *loss* and *crying* are associated with *sadness*

The related intensity lexicon (Mohammad et al.; Mohammad, 2018b) was created using best-worst-scaling where annotators were presented with several words simultaneously and asked to rank them from most to least associated with a specific emotion. The NRC Emotion Intensity Lexicon (a.k.a. NRC-EIL) is otherwise nearly identical to the NRC Emotion Lexicon in terms of the lexical items it contains. Seed annotations were also used as quality control; about 5% of the data was expert annotated and used to detect malicious or random annotations and when accuracy on the seeded annotations fell below 70%, all annotations by that annotator were discarded.

Examples from the Intensity Lexicon (Mohammad, 2018a):

- *outrage* (anger = 0.85, disgust = 0.47) is associated with a greater degree of *anger* than *irritate* (anger = 0.62)

[6] Available on `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`.

- *sobbing* (sadness = 0.69) is associated with more *sadness* than *sigh* (sadness = 0.40)

- *celebrated* (joy = 0.86, anticipation = 0.64) is associated with more *joy* than *humanity* (joy = 0.39, trust = 0.69)

The NRC Emotion Lexicon in particular is one of the most well-known and -cited emotion and sentiment resources, with a combined number of citations currently[7] totaling over 2000. This is likely due to the extensive work the creators took in selecting annotations and evaluating their dataset. The lexicon also has its own website which clearly and concisely presents the lexicon. It is simply very versatile, easy to understand and use, and the annotations are of very high quality.

3.4 Existing Annotated Datasets

Sentiment analysis can be done at different levels of granularity from word, phrase (e.g. news headlines), and sentence-level, to paragraph (or post/tweet) and document-level. Although the most common modality is text, sentiment analysis can be performed on at least images, video and audio, as well as biometric data. The classification can be of independent units or sequence labeling.

Fine-grained sentence-level sentiment analysis can increase the performance of many different types of applications (Pang and Lee, 2008), e.g. question and answering tasks (Wilson et al., 2009). Despite this, datasets annotated at sentence-level are relatively rare. This is likely due to the costs involved and the difficulty of annotating at this granularity. The more fine-grained the data is, the fewer sentiment and emotion clues are available to the annotator. The annotator may be forced to judge an entire sentence based on a single word (Boland et al., 2013). For successful sentiment analysis at the sentence-level, the inclusion of neutral sentiments is imperative according to Wilson et al. (2005) and Wilson et al. (2009) as it does not force the annotator to label a sentence as containing sentiment or emotions where there is none.

[7] Google Scholar, accessed September 17, 2020.

There are many different annotated datasets in existence, including for sentiment annotation. The most famous of these **sentiment datasets** are perhaps the Stanford Sentiment Treebank (Socher et al., 2013), the IMDB Movie Reviews dataset (Maas et al., 2011), the Sentiment140 dataset (Go et al., 2009), the Opin-Rank Review Dataset (Ganesan and Zhai, 2011), and the Amazon Product Data for sentiment analysis (McAuley et al., 2015; He and McAuley, 2016).

Bostan and Klinger (2018) analyzed 14 existing **emotion datasets**. Most of these datasets are multiclass, but single label[8]. AffectiveText (Strapparava and Mihalcea, 2007) and SSEC (Schuff et al., 2017) are the only multilabel datasets they have included in their study, suggesting multilabel datasets are rare. Virtually all of the analyzed datasets use some version of Ekman's (1971) universal emotions or Plutchik's basic emotions (Plutchik, 1980). This means that the number of classes is typically between 6-8. There is only one exception: CrowdFlower[9] with 14 categories. Other significant multiclass datasets are the SemEval 2018 task 1 subtask c dataset (Mohammad et al.) with 11 categories, and the GoEmotions dataset (Demszky et al., 2020) with 27 categories.

A large number of recent papers dealing with multilabel emotion classification use the SemEval 2018 dataset (Mohammad et al.), e.g. Ilić et al. (2018) and Jabreel and Moreno (2019). This dataset is based on tweets, as are many of the other datasets mentioned. This is likely due to a few reasons:

1. Twitter is popular and can be used to study a wide range of phenomena

2. Tweets are usually standalone by nature, meaning a lot of the relevant context is in the tweet itself.

3. The length of tweets are limited to 160 characters so the creator needs to be concise and clear, making it easier for an annotator to judge the content.

[8]Multiclass refers to classification schemes beyond binary, whereas multilabel refers to allowing a single data point to have more than one label.

[9]CrowdFlower was created in 2016 but has since been acquired by different companies at least twice and is now hard to find. It is currently owned by Appen.

4. Hashtags and emojis are common, aiding in judging the emotional tone and sentiment of a tweet.

5. Hashtags and emojis can also be used by themselves to automatically create labeled datasets with the hashtag as the label. As hashtags are labels assigned by the author, there should be no need to debate the accuracy of the label.

Reddit[10] comments, used by Demszky et al. (2020), are rarely self-contained as they are typically replies to posts or other comments. This means that context is not clear and comments are harder to judge. Emojis are also much rarer. Reddit comments are also typically longer than 160 characters and contain conflicting emotions, again making judgment harder. Reddit discussions take place in subreddits divided by topics of interest and content and style of speech varies greatly between different subreddits. This can be both a good thing and a bad thing. If the aim of the study is to compare different subgroups then the subreddits can be used to automatically group data points, however, the differences between subreddits can make the data as a whole less reliable.

In table 3.4 I list multiclass emotion datasets. The paper in which the dataset was described and evaluated is under (*study*), the domain of the source data (*source*), and (*cat*) shows the number of classes in the dataset. These datasets are explored further in section 4.3.

The datasets in table 4.3 can hardly be called similar. The domain, models, number of categories etc. vary. The accuracy scores are also heavily affected by the number of categories and the domain of the source data. It is much easier to achieve high accuracy scores when classifying fewer categories. This is true for both the machine learning model, as well as for the inter-annotator scores for the humans who originally annotated the dataset. I would like to note that the different sources and domains of the source data, are likely to lead to different levels of domain-dependence. Twitter datasets are likely quite transferable to analyzing other unseen Twitter data, but less likely to be successful in other domains. Whereas the datasets created by for example Tokuhisa et al. (2008) and Demszky et al. (2020) might perform reasonably well in domains outside of the original domain, but are perhaps not as accurate with in-domain data as Twitter datasets.

[10] `http://reddit.com`

study	source	cat
Strapparava and Mihalcea (2007)	news headlines	6+val
Liu et al. (2019)	books etc	8+neu
Schuff et al. (2017)	Twitter	8
Abdul-Mageed and Ungar (2017)	Twitter	6
Tokuhisa et al. (2008)	JP web corpus	10
Abdul-Mageed and Ungar (2017)	Twitter	24
Jabreel and Moreno (2019)	Twitter	11
Samy et al. (2018)	Twitter	11+neu
Yu et al. (2018)	Twitter	11
Huang et al. (2019)	Twitter	11
Demszky et al. (2020)	Reddit	27
Demszky et al. (2020)	Reddit	6
XED (English)	movie subtitles	8+neu

Table 3.4: Overview of emotion datasets

In the case of machine learning, the opinion is, rather unanimously, that domains should generally not be crossed as data trained on one domain might not work well applied to data from a different domain (Dave et al., 2003; Pang and Lee, 2008; He and Zhou, 2011). Therefore, a new annotated dataset is required for each domain, increasing the cost of using machine learning methods in practical applications.

3.5 Annotation Platforms and Other Tools

There are quite a few annotation platforms in existence, but most of them are optimized for transcription and even those that are not, have an annotation procedure that is not very suitable for emotion annotation. These systems typically have the user highlight a part of a text and then enter a category for the highlighted part (see e.g. brat by Stenetorp et al. 2012 or UBIAI[11]). Doccano (Nakayama et al., 2018) is an open source annotation tool which can be used for sentiment annotation as well, however, it too has a user interface

[11] `https://ubiai.tools/`

reliant on highlighting, rather than one allowing annotators to click on a button of their chosen class and automatically be presented with the next data point. IO Annotator has a simple and clean interface like this, but only allows for the annotation of *positive, negative* and *neutral*. Turksent (Eryigit et al., 2013) allows for more fine-grained annotation, but still only on one axis of valence. Additionally, it seems to be no longer available. Many, such as Tagtog[12], are not optimized for word or sentence-level annotation, and are instead focused on document-level.

Many of these annotation tools and platforms are certainly useful if the provided annotation scheme and annotation scheme design options match your needs. Some of these platforms offer additional tools such as bias detection (UBIAI) and automated inter-annotator agreement calculators, which could certainly prove useful. However, it seems that no platform offers a clean user-friendly interface for sentence-level emotion annotation. Therefore, we created our own, *Sentimentator*, which will be presented in section 5.2.

3.5.1 Parallel Corpora for Annotation Projection

A useful, if not necessary, tool for machine translation is parallel corpora where the same text is aligned to match the translated version of that text in another language. Such corpora can be useful in many different cross- or multilingual tasks, including in the study of emotion preservation in translation and annotation transfer through projection. As sentences are aligned a simple approach is to compare the emotion content of such aligned sentences in different language pairs. Here I briefly present two such corpora that I have used in my research.

Open Subtitles

Open Subtitles, OPUS, is a parallel corpus consisting of aligned subtitles from 60 different languages and 1,782 language pairs, originally created in 2016 (Lison and Tiedemann, 2016) and later updated in 2018 (Lison et al., 2019). The 2018 version contains 3.4 billion sentences and 22.2 billion tokens, and covers 208,000 movies and tv shows.

[12] `https://tagtog.net/`

29% of the subtitles are available in at least 10 languages (Lison et al., 2019). The subtitles in the corpus were extracted from the OpenSubtitles website.[13]

The translation quality varies widely between different subtitle-sets. It is likely that some of the translations have simply been translated using out-of-date machine translation such as Altavista Babelfish,[14] whereas others have been pet projects of professional translators. Regardless of the experience and devotion to quality of the translator, there are cultural and linguistic aspects that affect translation. Furthermore, the creators argue that subtitles are not direct translations, but should instead be viewed as "boiled down transcription" as translations compress the source material differently (Lison et al., 2019).

Although the OPUS subtitle corpus is carefully curated, there are some misalignments in the data as well as encoding issues. These consist of simple misalignments, often such that sentences do no align 1-to-1. Many such alignments have been corrected and can be matched using the IDs of the sentences, but some can not. There are also character encoding issues. In some cases it almost seems as a subpar OCR method has been used as the encodings suggest visually similar characters have been substituted for each other. In article VI we discuss these in more detail, but some common patterns we detected were:

- In many Finnish language documents, the character *l* was rendered as *I* (i.e. lower case "L" and upper case "i" respectively)
- In many Italian language documents, the character *à* was rendered as *ø*
- In many Italian language documents, the character *è* was rendered as *é*

The OPUS dataset comes with opustools (see e.g. Aulamo et al. 2020) that help users navigate and extract the data they need. In my research, opustools has been used to find aligned data for annotation projection and more information about the specific movies the annotated subtitles were associated with. Each line in the OPUS parallel corpus has its own ID that is linked to metadata such as language, release year, and the line placement in the movie as well as alignment.

[13] `https://www.opensubtitles.org/`

[14] The now defunct: `http://babelfish.altavista.com`.

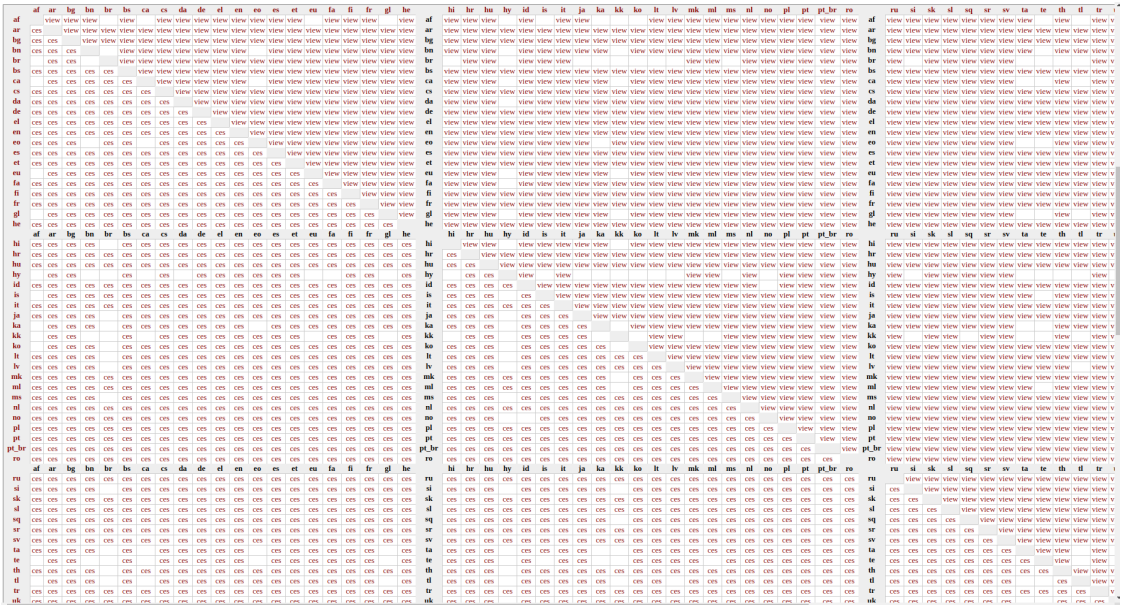

Figure 3.1: Screenshot of OpenSubtitles2018 interface[15]

Figure 3.1 illustrates how massively multilingual the OPUS corpus truly is.

The OpenSubtitles corpus as part of the OPUS[16] dataset should be viewed as a great tool in terms of the magnitude of data, and the applicability of the source data to various domains as it can be seen as a proxy for spoken language or even language used on social media (the topic is discussed in paper VII). However, one should keep in mind that the results may vary depending on the task due to the varying quality levels of both the translations themselves and alignments.

EuroParl

In paper V I use the EuroParl parallel corpus in conjunction with OpenSubtitles2018 and the NRC Emotion Lexicon. EuroParl is a parallel corpus of 11 languages with over 60 million words for each language. The texts have been retrieved from the European Parliament proceedings as published online. They have been widely used for statistical machine translation and other natural language processing tasks (Koehn, 2005). The corpus was updated in 2012 to include a further 10 languages. The advantage of Eu-

[15] `http://opus.nlpl.eu/OpenSubtitles-v2018.php`

[16] http://opus.nlpl.eu

roParl over OpenSubtitles2018 is the quality of the translations, as the translations have gone through professional accredited translators specializing in this particular domain. The corpus does, however, consist of far fewer language pairs and less data. The language is also more formal than that in OPUS and as these speeches in the parliamentary proceedings are very much scripted, they are even less representative of spontaneous natural language than movie subtitles that are also scripted, but scripted to represent spontaneous natural language.

Chapter 4

Sentiment Analysis and Emotion Detection

Sentiment analysis can be conducted by many different approaches. Typically, it is viewed as a classification problem. Sentiment analysis has progressed by leaps and bounds in the preceding decades and is nowadays usually done utilizing machine learning. However, lexical methods are still relevant and can be used independently or in conjunction with machine learning approaches. In this chapter I talk about the theoretical background of sentiment analysis and emotion detection. I start by discussing some of the core concepts and evaluation metrics. Then I talk about lexicon-based methods and move onto data-driven approaches. For both I also present related research. I conclude the chapter with a brief overview of practical downstream applications of sentiment analysis and a discussion on limitations.

4.1 Concepts and Evaluation

> "Given a set of of evaluative text documents D that contain opinions (or sentiments) about an object, sentiment analysis aims to extract attributes and components of the object that have been commented on in each document $d \in D$ and to figure out whether the comments are positive, negative, or neutral" (Liu, 2009, definition 1).

Sentiment analysis at the core is extracting subjective information from unstructured data, typically text. There are two main ways to do this, lexicon- or rule-based, and data-driven or machine learning. Hybrid methods, where both rules and machine learning are utilized exist too. Lexicon-based methods are cheaper to use, but in general not as reliable as machine learning approaches as they often disregard sequences and valence shifters such as negations and intensifiers. Such rule-based methods can of course be "upgraded" to take such things into account, but this often leads to very heavy and complex systems, and the benefits are typically quite small: Khoo and Johnkhan (2018) report an average increase in accuracy of approximately 0.025, that is 2.5 percentage points, between ignoring and including valence shifters. Machine learning methods on the other hand are good at learning from labeled data and are often able to correctly predict the label of unseen data. Such approaches require quality annotated data in order for the machine learning model to learn correctly. The drawback is the cost of the creation

of these annotated datasets, as well as the opacity of how the results were achieved (Jurek et al., 2015). They are however, very easy to evaluate quite conclusively. Even though lexicon-based approaches, even when optimized and fine-tuned, are outperformed by machine learning approaches (Kolchyna et al., 2015), there are advantages. Table 4.1 lists methods, as well as the advantages, and limitations of each approach.

Sentiment and Emotion Classification Approaches			
Type	Methods	Advantages	Limitations
Machine Learning	Multinomial Naive Bayes Multilayer Peceptron MaxEnt SVMs ELMO GPT LSTM BERT	High accuracy Learns from context	Heavy training procedures Costly annotations Mostly opaque
Lexicon-based	Matching Ensemble	Less domain-dependence Re-usability Transparency	Lower accuracy Word sequences are typically ignored

Table 4.1: Overview of machine learning and lexicon-based approaches

The most common modality for sentiment analysis is text, but multimodal approaches are becoming more and more common as social media increasingly moves from text to voice (e.g. WhatsApp), image (e.g. Instagram), and video (e.g. Snapchat) (Zadeh et al., 2017; Soleymani et al., 2017; Morency et al., 2011; Chen et al., 2017; Poria et al., 2015; Majumder et al., 2018; Rosas et al., 2013). There are three common research areas within multimodal sentiment analysis (Soleymani et al., 2017): spoken reviews and videos, Human - Computer Interaction (HCI), and computer vision: sentiment analysis of images.

Sentiment analysis is typically modeled as a classification task where a label from

a fixed set S is assigned to a given text. Often sentiment analysis tasks are modeled as binary classification tasks (*positive* or *negative*) and sometimes a third class, *neutral* is added. Emotion detection, sometimes called emotion analysis, is typically modeled as a more fine-grained multiclass classification task. Multiclass refers to classification schemes with more than two classes (technically binary + neutral is multiclass), and commonly there are at least six classes (see e.g. table 4.3). Most approaches use a single-label annotation scheme — that is, only one label per data point is allowed. Multilabel annotation schemes are more complex and allow a single data point to have more than one label as not all categories are mutually exclusive (cf. how the genre of a movie can be both *action* and *comedy*). The evaluation of multilabel classification can become very complex as the same methods that are used for single-label classification evaluation are often not applicable.

4.1.1 Evaluation of Results and Measuring Accuracy

To be able to compare, evaluate, and analyze the performance of models and quality of datasets, results need to be evaluated. The first part is obviously to split the data into training, testing, and possibly development and validation datasets. Core concepts in any classification or information retrieval tasks are accuracy, precision, and recall. Accuracy is the simple percentage of correctly classified data points:

$$accuracy = \frac{true\,positive + true\,negatives}{true\,positives + true\,negatives + false\,positives + false\,negatives}$$

Precision measures the percentage of correct positive data points out of all the samples that were predicted:

$$precision = \frac{true\,positive}{true\,positive + false\,positive}$$

Recall tells us the percentage of correctly classified positive data points:

$$recall = \frac{true\,positive}{true\,positive + false\,negative}$$

All of these measurements have their use, but for example, accuracy works best with balanced datasets. If we have a dataset where there is a significant majority class,

the accuracy percentage becomes rather meaningless. Consider the following situation: a dataset consists of tweets filtered using the hashtag *#depressed* and based on some additional criteria labeled as either *depressed* or not *depressed*. Considering the hashtag used to create the dataset we might find that 95% of the tweets are labeled *depressed*. This would mean that a model predicting all labels as *depressed* would have an accuracy score of 95% which does not necessarily indicate a good model. In such cases one solution would be to add more *non-depressed* tweets to the dataset in order to create balanced categories.

In many applications precision and recall are weighted differently as in some cases it is more important to minimize the number of false positives (e.g. spam and junk mail filters) and in other cases it is more important to find all positives and minimize false negatives (e.g. medical applications such as cancer screening and pregnancy tests). Usually increasing recall lowers precision and vice versa so it is important to consider which score is more important for different tasks.

In order to more accurately measure the performance of a model where both high precision and high recall are desirable, we can combine these measures into an f1-score. The f1 score is the harmonic mean of precision and recall:

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

The f1 score is always a value between 0 and 1, where 0 means that either precision or recall or both were 0, and 1 means that both precision and recall were perfect. In multiclass problems, precision, recall and F1-scores are defined per class. Macro and micro-averages can be computed to provide an overall quality estimation, where macro-averages are preferred for cases with large class imbalances (Narasimhan et al., 2016). Micro-averaging biases the score by the class frequencies whereas macro-averages treat each class as equally important.

The predictions of a classification are usually presented in a confusion matrix which shows the distribution of true labels and predicted labels as the number of true positives, true negatives, false positives, and false negatives (see table 4.2).

Confusion matrices for multiclass classification list all classes and offer a simple overview, often heatmap, of what misclassifications occurred. This means that it is easy

	Actual Positive	Actual negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Table 4.2: Simple confusion matrix

to see what classes are easily confused with each other. With multilabel classification, confusion matrices are not nearly as useful, as technically all categories could be true at the same time.

Measuring Annotation Agreement

κ scores are rarely used when evaluating the accuracy of machine learning models, however, it is probably the most widely used measure of inter-annotator agreement. κ is calculated as:

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

Where P_0 is the accuracy of the agreement between annotators and P_e is the expected frequency of chance. The value is between -1 and 1, where any score under 0 indicates that the agreement is less than chance, and a score of 1 indicates perfect agreement. Typically, a kappa score of 0.6 or higher is considered as adequate, and over 0.8 as nearly perfect (Landis and Koch, 1977).

However, Cohen's Kappa can be very misleading, and other methods which more accurately reflect the quality of the annotations should be considered in conjunction with κ scores (Delgado and Tibau Alberdi, 2019; Pontius Jr and Millones, 2011). Cohen himself warns that kappa can be specifically inadequate to measure such agreement when there is an imbalanced distribution of classes (Delgado and Tibau Alberdi, 2019), which is often the case with emotion-annotated datasets.

Delgado and Tibau Alberdi (2019) suggest that Matthews Correlation Coefficient (MCC) be used instead. In fact, Chicco and Jurman (2020) even suggest that it be used instead of f1 scores in binary classification tasks since it accounts for imbalance between the different sets. MCC is calculated as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where TP stands for true positive, FP for false positive, TN for true negative, and FN for false negative.

Perhaps the guidelines and expectations for which accuracy measure to use for evaluation will change in the coming years, but for now, the standard measurements in machine learning and sentiment analysis are still Cohen's κ for annotator agreement and macro-averaged f1 scores for model accuracy.

Loss Functions and Overfitting

Overfitting is when a machine learning model is bad at generalizing from training data to test data. This is why datasets are split into training and testing data and often a development, dev, or a validation set too. As the names imply, the training data is used to train the model, the test data to test the model, and the validation or development set is used to optimize or fine-tune the model. Simply put, if the model is significantly more accurate on the training data than the test/validation data, the model is likely overfitting, i.e. it works well on classifying the data it has already seen into the correct categories, but it does not generalize well and therefore is bad at classifying data it is seeing for the first time.

Loss functions, sometimes referred to as cost functions or error functions, are an integral part of machine learning and the evaluation of models throughout the learning process. Machine learning can be seen as an optimization problem with respect to the loss function, the so-called training objective. Simply put, a loss function is a measurement of the cost of inaccurate predictions and show us directly how well a model fits the data. Loss functions and validation data can aid us in choosing when to stop training a model and avoid overfitting. Figure 4.1 shows the relationship between loss and overfitting. When the training loss is going down, but the validation loss is rising, it is a clear sign of a model overfitting. The related concept of underfitting can also be a problem. This is when a model cannot generalize well with previously unseen data, but is not good at modeling the training data either.

There are many different types of loss functions and it is important to pick one that is suited for the model chosen and the type and structure of the data that is being used. There are two main types of loss functions: regression losses and classification losses

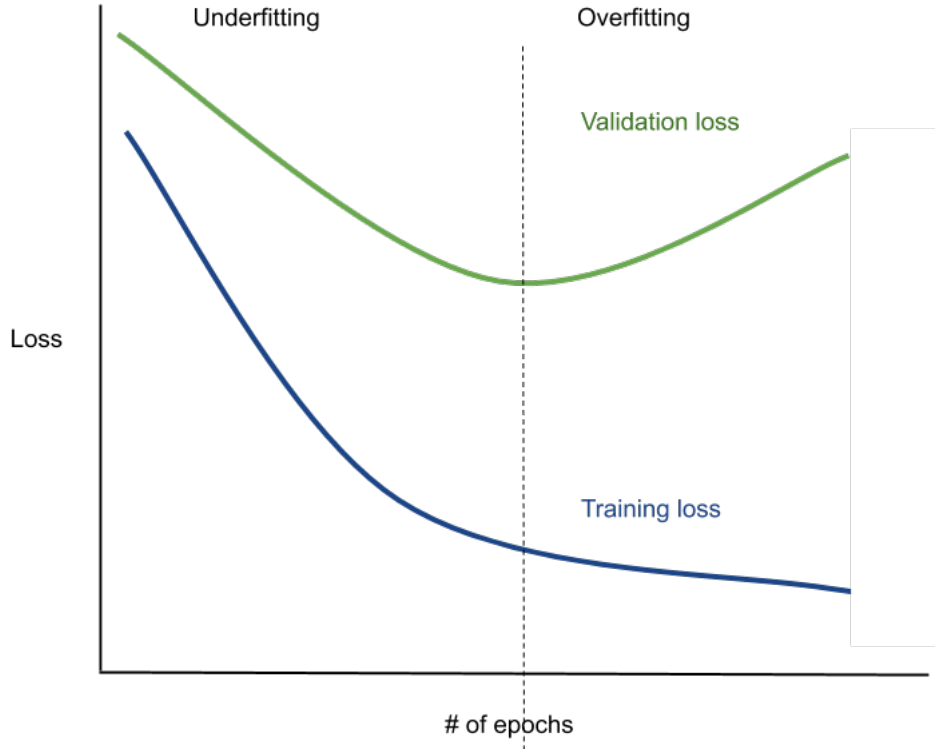

Figure 4.1: The relationship between loss functions and overfitting

(Bühlmann and Yu, 2003).

Classification loss functions are used with classification tasks. The most common classification loss functions are the log-loss, hinge-loss, and hamming loss. Log-loss is commonly used with multi-label classification:

$$-(Q \log(\hat{Q})) + (1 - Q) log(1 - \hat{Q}), Q \in 0, 1, \hat{Q} \in [0, 1]$$

Where Q represents the actual class, $\hat{Q}$ denotes the value predicted by the model, and $log(\hat{Q})$ is the probability of that class.

Hamming loss too, is used with multi-label classification:

$$\frac{1}{NB} \sum_{i=1}^{N} \sum_{j=1}^{B} \text{xor}(y_{i,j}, z_{i,j})$$

where $y_{i,j}$ is the target, and $z_{i,j}$ is the prediction. N is the total number of examples and B is the number of labels.

And Hinge loss, which is typically used with SVMs. Where $Q = \pm 1$, and L stands for loss:

$$L(Q, \hat{Q}) = max(0, 1 - Q \cdot \hat{Q})$$

Last but not least, binary cross entropy with logits loss used with multilabel classification in PyTorch[1]:

$$L_c(x, y) = l_c = \{L_{1,c}, ..., L_{N,c}\}^T ,$$

$$L_{n,c} = -w_{n,c} \left[p_c y_{n,c} \cdot log\sigma \left((x_{n,c}) + (1 - y_{n,c}) \cdot log \left(1 - \sigma \left(x_{n,c} \right) \right) \right) \right]$$

where L stands for loss, c is the class number; $c > 1$ for multi-label binary classification, $c = 1$ for single-label binary classification. n is the number of the sample in the batch and p_c is the weight of the positive answer for the class c. x is the input, i.e. the data input to be classified, y is the target, i.e. the label, and w is the weight (Paszke et al., 2019). σ represents the sigmoid layer as per the PyTorch documentation; "This loss combines a Sigmoid layer and the BCELoss in one single class. This version is more numerically stable than using a plain Sigmoid followed by a BCELoss as, by combining the operations into one layer, we take advantage of the log-sum-exp trick for numerical stability."[2] Cross-entropy loss functions are used for, for example, BERT-based multilabel classification approaches (Paszke et al., 2019; Demszky et al., 2020).

4.2 Lexicon-Based Methods

Lexicon-based approaches to sentiment analysis rely on sentiment lexicons to determine the sentiments of a text by comparing the sentiments of individual words or phrases. Lexicon-based methods are generally rule-based methods and often bag-of-words (BOW) approaches. A lexicon is used to assign scores, values, or labels to words in a text. These are then combined, summed or averaged, to obtain a final score at the sentence-, post-, or document-level. Other approaches, such as predicate calculus used by Al-Ayyoub et al. (2015) and methods combining lexicon-based approaches with machine learning (Zhang et al., 2011), have shown promise too when dealing with sentiments only.

[1] `https://pytorch.org/docs/master/generated/torch.nn.BCEWithLogitsLoss.html`
[2] `https://pytorch.org/docs/master/generated/torch.nn.BCEWithLogitsLoss.html`

A lexicon tends to have better scalability, i.e. they are more multi- and general purpose than most datasets for machine learning approaches as lexicons are typically generated from multi-domain dictionaries and datasets typically consist of data from a single source or domain. A lexicon can therefore more easily be used on multiple projects focusing on data from different domains (Taboada et al., 2011; Araque et al., 2019). Bandhakavi et al. (2017) however, show that domain-specific lexicons do outperform general purpose lexicons, and Medhat et al. (2014), rightly so, point out that general purpose lexicon-based methods might not be able to detect domain-specific orientations of emotion-associated words. Lexicons can be equally expensive to compile as annotated datasets, but since they are more domain-flexible, they can be re-used and many excellent emotion and sentiment lexicons already exist (e.g. the NRC Emotion Lexicon by Mohammad and Turney 2013).

Typically lexicon-based methods are used as simple rule-based lookup systems where each word in a dataset is compared to a lexicon. If the word is in the lexicon, a score is given according to the lexicon entry. Lexicon-based methods can also be combined with machine learning methods to create hybrid models. Hybrid methods have the benefit of being more accurate than just using lexicons and not as time-consuming or costly as machine learning methods. Kolchyna et al. (2015) use an ensemble method where they use a lexicon-based sentiment score as an input feature with a cost-sensitive SVM classifier to produce highly accurate predictions. Augustyniak et al. (2016) showed that lexicons can also be used as a first-stage classifier before employing a decision tree approach. They also used an approach they call “frequentiment” where the frequency of sentiment words were used as features in machine learning.

Purely lexicon-based approaches are nowadays usually limited to downstream applications where the creation of new annotated datasets is beyond the scope or budget of the study, and existing datasets are not suitable due to either annotation scheme or domain. Still, many have shown that lexicon-based approaches can be used successfully for, e.g. Twitter analysis focusing on sentiments (Musto et al., 2014; Jurek et al., 2015; Al-Ayyoub et al., 2015; Zhang et al., 2011) and other similar tasks where machine learning is commonly employed.

4.3 Data-driven Approaches

In this section I will mainly focus on supervised machine learning methods and different classification algorithms. Natural Language Processing often relies on machine learning approaches to teach computers to understand natural language[3]. Supervised machine learning is when you use labeled data to train a model and classify data into specific categories. The output is a category determined by the labels in the input data. Classifiers using supervised machine learning rely on two processes, the training of the model, i.e. learning, and the classification, i.e. predicting. Figure 4.2 illustrates this process.

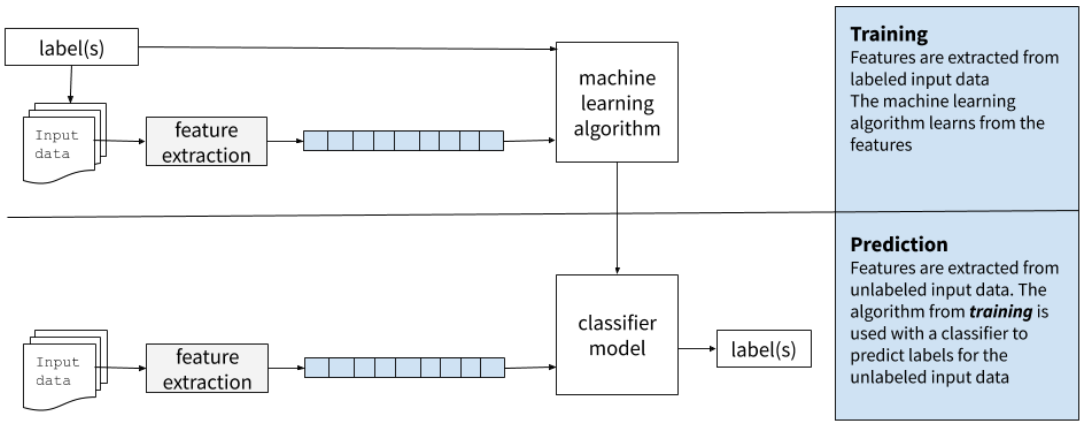

Figure 4.2: Supervised classification using a feature-based model[4]

The number of classes varies between tasks, but is always at least two. For sentiment analysis an example of a binary classification task would be classifying text into positive or negative. For most sentiment analysis tasks, the classification is at least ternary, including *neutral* as a category. For emotion detection the categories often range from 4 to 28 (Bostan and Klinger, 2018; Strapparava and Mihalcea, 2007; Demszky et al., 2020; Alm et al., 2005).

The task can be set up to only accept single labels or multiple labels per data point.

[3]Although computers are not actually taught to understand language as there is typically no link between form and meaning in machine learning (Bender and Koller, 2020), this is what the process is commonly referred to as.

[4]Adapted from the NLTK Book (Bird et al., 2009) chapter 6, figure 1.1.

The main difference is to instead of normal cross-entropy loss, to use binary cross-entropy with logits allowing the assignment of independent probabilities to each label. If a data point (word, sentence, paragraph, document, etc.) is only allowed to have one emotion label (e.g. only *anger* but not *disgust*) where multiple labels could be considered correct, it makes it easier to train the classifier and the results might even indicate a higher accuracy, however, in reality, many data points have multiple valid interpretations or reflect multiple emotions. The more categories that are allowed, the lower the expected accuracies of the evaluation of the model when dealing with classes such as emotions that are non-disjoint in nature.

One thing that all methods, models, and approaches have in common is the first step: pre-processing the data. Very rarely is data in such a format that it can be used as is. Different methods and different data require different pre-processing steps.

In most cases, the first step is going to be removing unnecessary characters. What these characters are depend on the dataset and the task at hand. This could be Twitter handles in Twitter data, or characters that appear in the data due to character encoding or platform incompatibilities. Sometimes punctuation and other similar characters are irrelevant to the task and needs to be removed. This can also be true for numbers and anything that is not a lexical part of the text.

Often stop words are excluded. Stop words can be any words that are deemed unnecessary for the task, but are typically common words that carry little lexical meaning such as conjunctions, some pronouns, common verbs (be, do, have), and prepositions. Stop word lists exist for many languages, but are often edited to suit the specific task and data. White space sometimes needs to be dealt with by reducing the number of consecutive white space or making the type of white space and line breaks used uniform across the text. This can be completed by Python's built-in functions, using packages like NLTK (Natural Language Toolkit) by Bird et al. (2009), or simply by using regular expressions.

Usually names of people, cities, businesses and such are not necessary, or it could be that including them would lower the accuracy of the classification task (Belainine et al., 2016) or it could even be illegal to include them due to e.g. privacy preservation. In this case, named entity recognition (NER) can be performed. NER recognizes entities such

as people, locations, corporations, organizations, etc. and these can then be generalized by substituting the found entities with generic tags such as <person> instead of “Sarah”, <location> instead of “Times Square”, or @USER instead of “YouTubefanatic342”.

Features can be selected manually or detected automatically. Features can be words, part-of-speech tags, bigrams, etc. Features correspond to the number of dimensions in the data. Text data tends to be high-dimensional and complex, and since for every feature, n we get 2^n feature sets or feature vectors, reducing features can remarkably improve training time and reduce overfitting (Bermingham et al., 2015).

Often, datasets are split 70-20-10 into training, validation or development, and testing sets respectively. The labels are generally human-annotated, but can also be automatically annotated using some feature in the text such as */s* in Reddit posts (Khodak et al., 2017) or the hashtags *#sarcasm* and *#not* on Twitter posts (González-Ibánez et al., 2011; Sulis et al., 2016; Poria et al., 2016) to label certain posts as sarcasm. Using hashtags as labels as is, is also possible (Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017). With small datasets, the splitting of scarce training data into an additional validation or development sets might not be feasible and it might be preferable to use a 90-10 split into training and testing and/or use cross-validation. Cross-validation, or k-fold cross-validation, randomly divides a dataset into k groups or subsets of similar size, where k is the total number of folds or groups. The first fold becomes the test set and the remaining $k - 1$ folds are used for training. Therefore, each subset is used once as a test or validation set, and all subsets are used $k - 1$ times for training, with the total number of training instances are equal to k (see e.g. James et al. 2013).

It is often useful to have a benchmark using a simpler model or in case of very imbalanced datasets assume that everything is classified into the majority dataset as a benchmark accuracy. This means that in a dataset of 100 data points where 90 are of type A and the rest of type B, the baseline accuracy when everything is classified as type A would be 90%, and therefore the desired accuracy of any model should be higher than the 90% of the baseline results.

There are many popular machine learning models used for different tasks. For a single task, often several different models are used in order to see what model yields the best scores for that specific task. Common models used in sentiment analysis are SVMs

(Support Vector Machines), CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), LSTMs (Long-Short Term Memory), and BERT (Bidirectional Encoder Representations from Transformers). In the following sections I first explain machine learning models in general before looking at specific approaches used in sentiment analysis.

4.3.1 Classic Classification Models

MLP, NB, MaxEnt

A multilayer perceptron is one of the simplest classifiers for machine learning. It consists of at least three layers[5] (see figure 4.3) of which at least one is hidden (Pal and Mitra, 1992). Multilayer perceptrons were very popular in the early days of computational classification tasks, but lost in popularity when other equally simple but better classifiers were developed.

Naïve Bayes Classifiers are another simple type of classifiers that rely on conditional probability models (Webb et al., 2005). This means that a Naïve Bayes classifier learns the probabilities of feature co-occurence from the labeled training data and assigns the data its most probable label. Although most modern models outperform Naïve Bayes classifiers, they are still used for their simplicity (Zhang, 2005; Liu, 2012).

Maximum Entropy (MaxEnt) classifiers, also known as multinomial linear regression classifiers, relaxes the independence assumptions that Naive Bayes makes and computes feature weights in an iterative optimization procedure. MaxEnt classifiers are known to produce more accurate results but the training procedure is more expensive. In many cases, a simple Naive Bayes model constitutes a quite competitive baseline and in our experiments we skip the application of the heavier MaxEnt model but instead pro-

[5] "A layer is the highest-level building block in deep learning. A layer is a container that usually receives weighted input, transforms it with a set of mostly non-linear functions and then passes these values as output to the next layer. A layer is usually uniform, that is it only contains one type of activation function, pooling, convolution etc. so that it can be easily compared to other parts of the network. The first and last layers in a network are called input and output layers, respectively, and all layers in between are called hidden layers." (Dettmers, 2015, 1).

[6] Image inspired by: `https://medium.com/pankajmathur/a-simple-multilayer-perceptron-with-tensorflow-3effe7bf3466`.

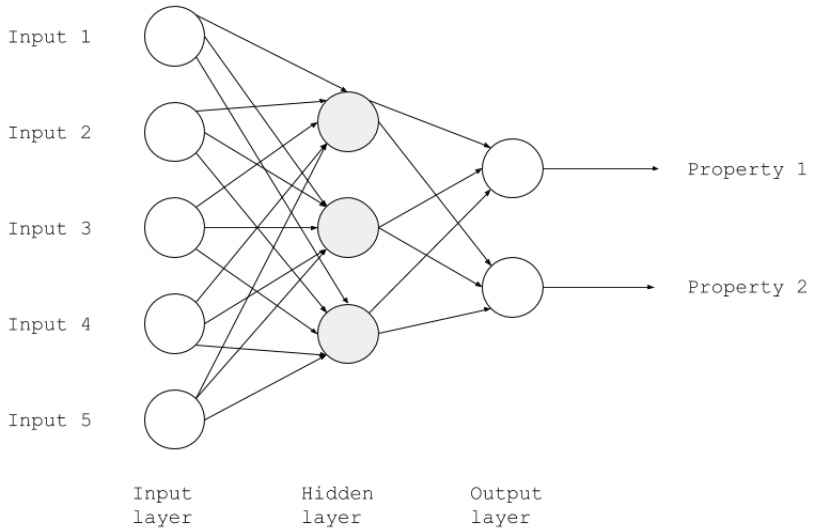

Figure 4.3: Multilayer Perceptron[6]

vide another strong feature-based classifier based on support vector machines (SVMs) introduced below.

SVMs

SVMs are non-probabilistic binary linear classifiers closely related to logistic regression. As the name implies, SVMs utilize vectors in feature space and try to find the optimal hyperplane that separates classes (see fig. 4.4). The number of dimensions is equal to the number of features. This hyperplane search can be extended to patterns that are not linearly separable by transformations of original data to map into new space where the data can be separated using a simple line. This is known as the Kernel function (Cortes and Vapnik, 1995; Awad and Khanna, 2015; Berwick, 2003). SVMs work best with fewer classes that do not overlap (Joachims, 1998).

The benefits of SVMs include their high accuracy when dealing with few clear categories and strong dependencies between features (Sculley and Wachman, 2007). They are highly memory efficient as the classification only relies on the support vectors and all other data is ignored. However, SVM training takes a long time and training time

[7]Source of original figure: `https://cg2010studio.com/` CC BY 3.0 TW.

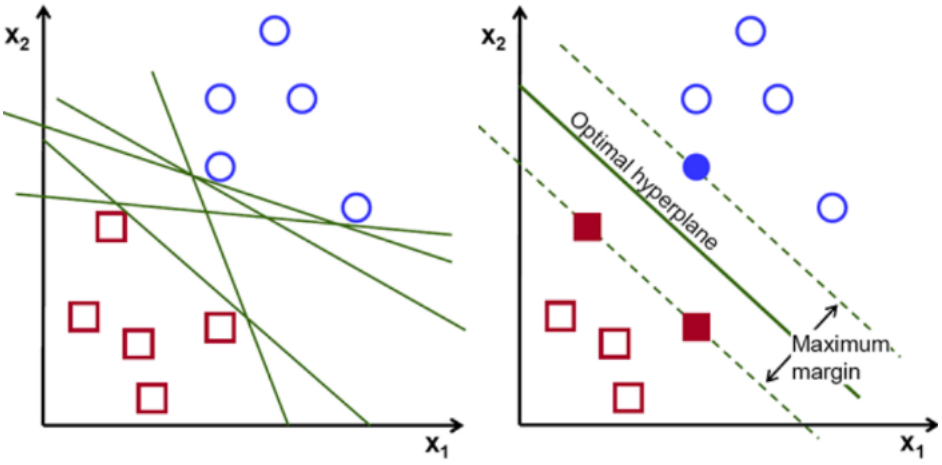

Figure 4.4: Support Vector Machine hyperplane[7]

grows exponentially with the data and number of features (Joachims, 2006; Scholkopf and Smola, 2018). As SVMs rely on the estimation of a separating hyperplane between two classes they are basically designed for binary classification tasks. It is possible to extend SVMs for non-binary cases by training multiple SVMs as binary models and combining these binary models for multiclass classification. This can be done as either one-vs-one classification or one-vs-rest (Mathur and Foody, 2008):

- In one-vs-one classifiers, where n is the number of classes $\frac{n*(n-1)}{2}$ is the number of classifiers needed for multiclass classification as all classes are classified as binary against all other classes separately.
- As the name implies, in one-vs-rest classification the data is split into one class at a time versus all the other classes so the number of required classifiers is n where n is the number of classes.

4.3.2 Neural Networks

Neural Networks are machine learning models that mimic how the human brain works. They consist of interconnected nodes that work similarly to the neurons in our brain. To replicate the connections of neurons, artificial neural networks use positive and negative weights to model excitatory and inhibitory neural responses.

Abdul-Mageed and Ungar (2017) use distant supervision to create the labels for their EmoNet dataset. In their case this means 665 Twitter hashtags that are grouped as subsets of 24 emotions: 3 intensity levels of Plutchik's 8 basic emotions. Their approach is single label and they outright state that their goal is the creation of non-overlapping categories.

They obtain some remarkable accuracies using GRNNs (Gated Recurrent Neural Nets): for all 24 classes they achieve f-scores of 87.47% and for predicting the 8 primary dimensions (*serenity, joy, ecstasy* are all mapped to *joy*) they receive an f-score of 95.68% which is truly astonishing. They attribute their success to the magnitude of training data: 1.6 million Tweets. I agree that it is very likely a part of the reason for their high accuracies, but I suspect the bigger reason is the fact that they managed to create truly disjoint categories of emotions. It would be very interesting to examine the data to see how they managed to create such separate categories based on simply hashtags, but unfortunately neither the data or the code seems to have been made available.

Another option is to apply transfer learning and a classifer that is built on top of a general-purpose neural language model. BERT is an open-source deep pre-trained language model developed by Google (Devlin et al., 2019) that can be employed in such a way. Judging from recent SemEval submissions (Basile et al., 2019; Zampieri et al., 2019b, 2020), BERT has almost completely taken over as the most widely used model in NLP downstream tasks despite being released as recently as 2018. It is easy to see why, almost all high-performing systems in these SemEval tasks used BERT or a combination of BERT and other models. This is in line with its results in natural language understanding tasks such as GLUE, SQuAD, and SWAG (Devlin et al., 2019).

BERT's high performance, in comparison to prior competitors, is a result of a few different things: (1) BERT is bidirectional. This means that BERT fully utilizes the information in text sequences from both left-to-right and right-to-left unlike other models that are either or (such as GPT Radford et al. 2019) or concatenate (such as ELMo by Peters et al. 2018). This means that every instance of a word in BERT has its own vector and it takes the entire context of the word into account, that is both what comes before and what comes after. (2) It uses both Masked Language Modeling (15% of words in the training data are masked and the network tries to predict the masked words based on the context provided by the non-masked words) and Next Sentence Prediction (are two

given sentences coherent together based on training corpora of consecutive sentences and random non-consecutive sentences). (3) It was trained on two massive corpora: the BooksCorpus and Wikipedia using the training objectives from (2) (Devlin et al., 2019).

BERT comes in two different versions "out of the box", *base* and *large*. The *base* version was trained on the BooksCorpus (800M words) (Zhu et al., 2015) and the *large* version was trained on Wikipedia (2,500M words). There are also official English, Chinese, and multilingual models from Google. Other language-specific models exist, such as the Nordic BERT model, which was trained by a company called botxo[8]. Usually these language-specific models have been trained by third parties, such as researchers or private companies, and are often, but not always, open source.

mBERT, Google's multilingual BERT model, has been shown to work well with all the 104 languages it covers (Wu and Dredze, 2019; Libovický et al., 2019; Wu and Dredze, 2020), and outperforms the crosslingual XNLI (Conneau et al., 2018) model easily (Wu and Dredze, 2019). However, there are differences between the languages. Wu and Dredze (2020) cautions against using mBERT alone with the lowest resource languages (bottom 30th percentile) or the top 10%. They recommend training on pairs of closely related languages, which is precisely what botxo did with Nordic BERT.

FinBERT was pre-trained on over 3 billion tokens of Finnish text from different news corpora (YLE, STT), online discussion corpora (Suomi24), and internet crawls. The amount of Finnish texts that were used to train mBERT, is approximately 3% of the amount used to train FinBERT. This makes FinBERT able to outperform not only mBERT, but also previous Finnish-specific models (Virtanen et al., 2019).

Demszky et al. (2020) use the BERT-base model to evaluate their data adding a cross-entropy loss function to support multilabel classification. They receive f1 scores of 0.46 for their 27 classes. They further test their BERT model trained on the GoEmotions dataset on other benchmark datasets and show that it yields good results in domains outside of Reddit too. BERT and similar models are indeed powerful new tools that are likely to stay in the field of language technology.

GoEmotions by Demszky et al. (2020) consists of Reddit comments that have been manually reviewed to reduce profanities, "adult tokens", offensive words targeted at any

[8] `https://github.com/botxo/nordic_bert`

particular ethnicity, gender, or sexual orientation[9] and remove identifying information. They also balance sentiments and emotions by downsampling weakly-labeled data, and balance subreddits and comment lengths. These comments are then manually annotated by human annotators into 27 emotion classes plus *neutral*. The annotators were native-English speakers from India, which means that although their linguistic background is similar, their cultural background is quite different and thus they might annotate the data created by mostly Americans differently to how someone with a more similar cultural background might.

In table 4.3 I have compiled some of the most relevant emotion datasets that at least in some respect are similar to my work and use some type of neural network. I list the paper in which the dataset was described and evaluated (*study*), the domain of the source data (*source*), and the most successful machine learning model (*model*) used along with the accuracy score (*macro f1*) achieved with said model, and the number of classes or categories (*cat*) that were used as labels. If the data was annotated as multilabel, I have marked it as *multi*. Some papers only reported a micro f1 and no macro f1 score. These scores have been marked with a μ.

study	source	model	cat	multi	macro f1	accuracy
Strapparava and Mihalcea (2007)	news headlines	“CLaC system”	6+val	yes	-	55.1%
Liu et al. (2019)	books etc	BERT	8+neu	no	60.4%μ	-
Schuff et al. (2017)	Twitter	BiLSTM	8	yes	N/A	N/A
Abdul-Mageed and Ungar (2017)	Twitter	GRNNs	6	no	N/A	95.7%
Tokuhisa et al. (2008)	JP web corpus	k-NN	10	no	N/A	N/A
Abdul-Mageed and Ungar (2017)	Twitter	GRNNs	24	no	87.5%	N/A
Jabreel and Moreno (2019)	Twitter	BiRNN/GRU	11	yes	56.4%	59%
Samy et al. (2018)	Twitter	C-GRU	11+neu	yes	64.8%	53.2%
Yu et al. (2018)	Twitter	DATN	11	yes	44.4%	45.7%
Huang et al. (2019)	Twitter	BiLSTM/ELMO	11	yes	70.9%μ	59.2%
Demszky et al. (2020)	Reddit	BERT	27	yes	46%	N/A
Demszky et al. (2020)	Reddit	BERT	6	no?	64%	N/A
XED (English)	movie subtitles	BERT	8+neu	yes	53.6%	54.4%

Table 4.3: Overview of emotion datasets

[9] It is unclear why they do this as such information has been shown to help identify negative emotions, and could be very valuable in downstream applications dealing with offensive speech.

As the examination of these datasets and EmoNet (Abdul-Mageed and Ungar, 2017) and GoEmotions (Demszky et al., 2020) in particular show, it is very difficult to create emotion datasets. It is even harder to compare such datasets as so many factors affect the evaluation scores and the annotations. As discussed in previous chapters, the annotation task is inherently subjective with emotion annotations as the categories are not necessarily independent. Many different models have managed to obtain impressive accuracies when the model has been tailored to a specific dataset, but the same model is not necessarily as accurate on a different dataset.

4.4 Practical Applications

Traditionally sentiment analysis has been used in the private sector to make stock market predictions or to follow how a specific brand is being discussed in the public. In academic research, the focus is usually more on either optimizing algorithms, models, and systems for improved accuracies and predictive power, or on applying sentiment analysis to practical research. This research can be e.g. medical, or humanities or social sciences related.

Offensive speech and hate speech detection are fields of research closely linked with sentiment analysis and emotion detection. The approach is generally very similar to binary or ternary sentiment analysis where a dataset uses the labels offensive and non-offensive (Zampieri et al., 2019b; Sigurbergsson and Derczynski, 2020), no hate, weak hate, and strong hate (Del Vigna et al., 2017), or something similar and the model is trained using this data to correctly predict whether unseen data points are offensive or hateful. Some researchers (Schmidt and Wiegand, 2017) have tried to combine sentiment analysis with hate speech detection, but almost all attempts have been practical in nature and yielded very little theoretical knowledge on the relation between hate speech and emotions.

Politics and sentiment analysis have been linked in research in many different ways. Sentiment analysis has been used to predict election results (Bermingham and Smeaton, 2011; Ramteke et al., 2016), to examine the emotional rhetoric of political communication (Haselmayer and Jenny, 2017; Mohammad et al., 2015), and to predict political

party identification (Biessmann, 2016). In both political science and communication studies, the rhetoric of populist parties has been identified to be both more emotional and to elicit more emotional reactions from the public (Wirz, 2018). Most sentiment analysis studies conducted by political science scholars rely on lexical methods (see e.g. Crabtree et al. 2018 and Puschmann and Powell 2018), and it seems a specific program, called LIWC (Pennebaker et al., 2003) is used in many of these cases. This program uses simple lexical lookup to calculate the percentage of emotion-words of different types in the input data, but also presents information such as percentage of specific part-of-speech tags. This program is criticized by many (Danescu-Niculescu-Mizil et al., 2011; González-Bailón and Paltoglou, 2015; Schwartz and Ungar, 2015; Su et al., 2017; Puschmann and Powell, 2018) for, e.g. conflating words with discrete units of meaning and over-generalization and application of psychometry to written language (Puschmann and Powell, 2018). Furthermore, it could ne argued that it is a black-box tool, meaning it is difficult to know exactly why it arrives at particular solutions in different cases, making users blindly trust the results without the ability to critically examine them.

Although some of the criticism towards LIWC also applies to other lexicon-based methods, the problems specifically with LIWC is partly because of its widespread use in traditional fields where the understanding of the more technical aspects of the program and lexicon-based methods is likely to be more limited. Unlike most sentiment and emotion lexicons, LIWC is not free. And even though it is possible to extract the dictionary from the software once purchased[10] if one possesses sufficient technical expertise making it technically not a black box tool, I would still claim it is often used as one. Nonetheless, LIWC is one of the oldest and most continuously updated sentiment "dictionaries" available for multiple languages even if it is virtually unknown in the fields of computational linguistics and language technology.

In literature studies, affect has long been a topic of interest. In most cases though, the approach has been purely qualitative and relied on close reading and human interpretation of the material (Rossi, 2020; Hollsten and Helle, 2016; Isomaa, 2012; Van Lissa et al., 2018; Sklar, 2013; Menninghaus et al., 2017; Keen, 2007; Lyytikäinen, 2017; Ngai,

[10]At least in the older versions before 1.6.0. according to `https://github.com/chbrown/liwc-python/issues/5`.

2005). The syuzhet package for R (Jockers, 2015b, 2017) is a well-known tool, especially in the field of digital humanities, for analyzing the plot structure of novels. The theory behind it is simple: shifts in emotional valence (positive-negative) can be used as a proxy for plot movement. The package has had its fair share of criticism, technical regarding the use of a low-pass filter as a smoothing function (Swafford, 2015, 2016), and theoretical regarding the validity of using the macro-level emotion and sentiment distribution as a proxy for plot movement. Jockers responds to this criticism (Jockers, 2015a) and shows how the criticism is not fully warranted. He says:

> " The point is to reveal a latent emotional trajectory that represents the general sense of the novel's plot...If the overall shape mirrors our sense of the novel's plot, then the tool is working...[*syuzhet*] successfully reveals the overall shape of the narrative." (Jockers, 2015a)

I would add to this that, yes, the tool is not perfect, and individual words do not necessarily even accurately capture the emotions or sentiments in even a single sentence, however, when combined to paragraph, page, and chapter-level analyses, *syuzhet* has been shown to quite accurately reveal the important shifts in the narrative trajectory of a novel.

The reliability of the package has been evaluated by comparing human annotations of valence to the plots produced by the syuzhet package and seem to line up well in general (Jockers, 2016). Jockers (2017) also mentions that the package with its built-in support for multiple languages seems to work well with non-English literary works too, but that this is not heavily tested. Although this tools is well known and widely referenced, there are no academic studies that use it to investigate plot arcs, structures, or syuzhet[11] of novels, although one study looks at it from a methodological point of view (Gao et al., 2016) and another one uses syuzhet to analyze post-emergency and training exercise debrief reports (Gunawan and Aldridge, 2018).

I think there are a few reasons for the lack of studies using syuzhet in literary analyses, and they mostly have to do with the problem of who the target audience would be

[11] In literary theory, particularly Russian formalism, *syuzhet* is the term used to describe how a novel is narrated, and contrasted with *fabula*, which is the chronological order of events (Gorman, 2018). In this dissertation, syuzhet refers to the R-package unless otherwise noted.

for such studies. Those who would have been most likely to both hear about *syuzhet* and want to use it are probably digital humanities scholars focusing on computational literary analysis, and for this group there are plenty of people eager to use such tools, but with limited levels of technical expertise (Wilkens, 2015; Landt, 2010; van Es et al., 2018). However, the early criticism of *syuzhet* made many wary of using it at all, including this aforementioined group. Traditional literary scholars would not have heard of it unless they were already interested in computational methods for literature, and computer scientist, even those working in computational humanities, would have had no use for it as any more serious project with *syuzhet* would still have involved qualitative analyses as well and the development of new quantitative approaches. Even those who would know how to use *syuzhet* correctly within its limitations, are unlikely to use it since it does not deal with aspects relevant to their research. *Syuzhet* is probably most relevant to those doing traditional research into affect in literature and wanting an objective tool as guide to their subjective interpretations.

There are, however, scholars who investigate emotions in literary works with quantitative and computational methods. Mohammad and Turney (2013) tracked emotions in both novels and fairy tales with a focus on visualization. They showed that fairy tales have much wider densities and distributions of emotions than novels do. Alm and Sproat (2005) investigated the emotional temporal plot structure and emotional distributions in Grimm's fairy tales with a focus on emotional sequencing and positioning. They wanted to find out if fairy tales more frequently begin neutral and end happy (they do to a significant degree), if neutral dominates emotional sentences' immediate context (it does except for *disgust*), and which emotions are more often prolonged, and which are not (surprise or fear are short term emotions compared to sadness and anger).

Kim and Klinger (2018) present a fairly thorough survey on sentiment and emotion analysis for computational literary studies. They list at least five top categories with further sub-categories of different types of sentiment and emotion analyses. The top categories they use are:

- Emotion Classification
- Genre and Story-type Classification

- Temporal Change of Sentiment
- Character Network Analysis and Relationship Extraction
- Other types such as Emotion Flow

A clear majority of the studies they mention use dictionaries and rule-based approaches.

4.5 Shortcomings, Drawbacks & Challenges

Deep learning has had a massive impact on all fields of natural language processing and understanding, including sentiment analysis, but the reverse is also true. Henderson (2020) writes that "the immense success of adapting deep learning architectures to fit with our computational-linguistic understanding of the nature of language will doubtless continue, with greater insights for both natural language processing and machine learning." Indeed, we are still far from having solved natural language understanding (NLU), however, it is probably disingenuous to some degree to claim that the issue is simply with NLU.

Puschmann and Powell (2018) write that conflating scoring and interpretation into the same process is a common misunderstanding about what sentiment analysis is. They suggest that sentiment analysis is simply scoring words and the NLU is done by the researcher interpreting the scores. This is certainly true in lexicon-based methods, and in some cases machine learning as well. As discussed earlier, it is of utmost importance to say only what your data shows. By this I mean that if you are using a lexicon to match words in a text, you can only claim things about the emotion word distribution. You can not, for example, claim anything about the mindset or even intents of the author. Sentiment analysis is a wonderful quantitative tool for many types of text analysis, but at the very least a thorough understanding of both sentiment analysis methods, and the source data needs to accompany the analysis.

Supervised machine learning approaches require annotated training data. This type of data is expensive to produce and the availability of such datasets are limited. Such annotated datasets are often limited by annotation quality and the simple fact that there

is rarely a single objectively correct annotation for a word, phrase, sentence, paragraph, or document (Mohammad, 2016; Kiritchenko and Mohammad, 2017, 2016), but also quantity, as machine learning models require vast amounts of data to learn properly.

Unsupervised machine learning methods try to overcome this limitation by trying to "figure out" the data by using different algorithms. These often fail to achieve similar accuracies as supervised approaches (Cheng et al., 2017; Scheible and Schütze, 2012) and cannot provide explicit labels, but are preferable in some cases because they can be much cheaper to use than the datasets for supervised learning.

Some of the most common challenges in sentiment analysis are data-imbalance and sparsity, domain and language dependence, and the cost and noise of annotations (see e.g. Alm 2012 for a more thorough exploration of these challenges). Some of these challenges can be dealt with to some extent, but it is often time-consuming and expensive. Sentiment analysis has traditionally focused on two areas of inquiry: 1) Computer science and computational linguistics, and 2) marketing and business. When using sentiment analysis for practical applications outside the typical division in particular, one needs to carefully consider what is meant by sentiments and emotions and what exactly the methods used are picking up in the data.

When creating annotated datasets or lexicons, it is important to inform the annotators of how they are to annotate: from whose point of view? With or without context? But even with perfect annotators producing perfect annotations with high inter-rater agreement scores, the text that is being analyzed can be noisy and full of abbreviations or spelling mistakes, or the writer might use anaphora, sarcasm, or irony that the algorithm or model is unable to recognize. If using translated texts at any step in the process of creating or analyzing the data, what is the translation quality and stylistic choices if a human translator has been used? Emotions are fairly consistent across different cultures, but they way they are typically expressed in text are not. Mostly these datasets are created with English as a starting point, and models tend to be optimized for English.

All of the above are things that need be given serious consideration when designing datasets and choosing the method of analysis. None of them are properly solved yet, although improvements and progress are being made at a consistent pace. This dissertation aims to contribute to solving these questions.

Chapter 5

Contributions

The methods and approaches described in the previous sections are the ones that I have used in my research. Influential studies in sentiment analysis and emotion detection typically use similar or comparable methods. A majority of these papers, in contrast to my research, are heavily focused on improving machine learning methods for sentiment analysis. My focus has been more on utilizing these methods for better applicability for related research such as the creation of datasets for low-resource languages and practical downstream applications.

5.1 Improved Lexical Resources

Like many, I started my doctoral studies with a literature overview. There were plenty of Twitter-based datasets that were discussed in the literature, and many more studies that dealt with sentiment analysis on a two-dimensional axis of positive and negative polarities. The few studies that discussed emotion detection did so from a model-improvement point of view, i.e. they were not terribly interested in the data itself, rather they used the data as a resource to evaluate their model.

One thing nearly all studies had in common was that the language of the dataset was English. This is when I came across the NRC Emotion Lexicon. Since I wanted to produce sentiment and emotion resources for Finnish and other, even more low-resource languages, I used the OPUS subtitle and EuroParl parallel corpora to evaluate the preservation of sentiments and emotions in translation and simultaneously evaluate the translations in the lexicon too. This study made it clear that a better curated lexicon for Finnish was necessary to produce high-quality lexicon-based emotion detection. This is how SELF and FEIL came into existence (paper I).

In papers V, IX, and X I use mainly or at least partially lexicon-based methods. These methods range from simple lexical lookups to lexicon-based vectors and using my own lexicon with *syuzhet*. Except for paper V the lexicons used are the SELF and FEIL lexicons presented in paper I. The lexicon used in V is the original 2016 version of the NRC Emotion Lexicon which was augmented by re-translating all the words with Google Translate — the same step as later taken by the creators themselves (in 2016 there were only about 6,000 Finnish words in the official NRC lexicon, this was updated

in 2017 to the full 14,182 words by using the improved Google Translate).

5.1.1 Emotion Lexicons for Finnish: SELF and FEIL

I created the SELF and FEIL lexicons based on the NRC Emotion Lexicon and the NRC Emotion Intensity Lexicon. I describe the process in paper I. The NRC Emotion Lexicon (a.k.a. EmoLex) was originally created by annotating English data. It has been translated by its creators into a multitude of languages using Google Translate. These translations do not cover all the English words; out of the 14,182 terms in the NRC Emotion Lexicon, only about 6,000 had been automatically translated into Finnish in the original multilingual lexicon. A new version with updated translations was published in 2017. This newer version had more translations from the improved coverage of Google Translate, but suffered from most of the same problems the earlier lexicon did.

Based on the automatically translated version of the NRC Emotion Lexicon and the Intensity Lexicon for Finnish, I analyzed and revised the translations, making substitutions, additions, and deletions where necessary. The original translations had a large percentage of duplicates due to the fact that Google Translate chooses the most common translation/word as the correct translation. This means that words that are near-synonyms in the source language are all translated into the same word in the target language.

As a first step, Google Translate was used to re-translate the remaining 14,000+ words in the NRC Emotion Lexicon. Google Translate chooses the most common translation of a word, which in this case means that nuances are lost and there are more duplicates than unique words in the resulting lexicon. If duplicates are not removed, common words are heavily weighted and the results are skewed, although this varies depending on the chosen methodology.

The duplicates were marked and carefully examined to find better translations where applicable and removed in other instances where no other appropriate translation was possible. In some cases both the British and American spellings of words were present in the lexicon, e.g. *grey/gray*, *tumor/tumour*, and were therefore removed as duplicates. There were two different common cases of over-generalization;

1. Where one word had been unnecessarily generalized: e.g. the word *emaciated* was

translated as *laihtunut* which means “to have lost weight”, but a better translation in terms of connotation and intensity would be *riutunut* which is near-identical in meaning and connotation with *emaciated*.

2. Where several near-synonyms in the source language had been all translated into one or two words in the target language. Case in point: *godly, pious,...* (see table 5.1)

Original English	Google Translate	Edited Translation
pious	hurskas	hurskas
devout	hurskas	harras
saintly	hurskas	pyhimysmäinen
godly	hurskas	jumalinen

Table 5.1: Example of lexicon editing: the case of *hurskas*

These findings are in line with the mistranslations found by Salameh et al. (2015) and Mohammad et al. (2016) described in section 5.3.

A related issue occurred with words that in English have slightly different meanings, but only one word for them in Finnish. E.g. Finnish does not differentiate between poison-venom-toxin. It is all the same word *myrkky*.

Another problem with certain words is the semantic shift that has taken place. The words have different connotations in modern society than they had 80 or even 50 years ago. This is a delimitation for any study using texts older than a few decades where the semantic properties of words are relevant, regardless of method. Societal and cultural differences between the US and Canada on one hand, and Finland on the other also affect the relevance of the annotations. This is apparent in, for example, the emotional connotations regarding religious words.

It is not always necessarily the best translation that is the best choice either. A translation which is not very close, but has similar connotations and, most importantly, does not yet exist in the lexicon, can sometimes be the better choice due to the way the lexicon matches words in the text. Especially very positive or negative words have a lot of synonyms or near-synonyms with all of them translated into the same word in Finnish.

Without careful examination and editing of the lexicon, this would have a serious impact on the reliability of the results as it would multiply the emotion scores and therefore misrepresent the data.

For some translations where in English the word form was interpretable as either/both a noun and a verb, the entry was duplicated and a second Finnish translation was added for both the act itself and the performing of that act. In some cases, the translation seemed fine at first glance, but when the related emotions and intensities were examined more closely, it was evident that this was not the case. Another common issue was that the Finnish translation was too general where the English word was very specific. In some cases it was easiest to remove entries that were ambiguous or unlikely to be found in the chosen works of literature such as modern slang words. In some cases, the correctness of the translation was secondary to the similarity of the corrected form to the emotions and intensities. Here are a few examples of actual changes that were made:

Original English	Automatic Translation	Corrected Form(s)
birch	koivu	-entry removed-
emaciated	laihtunut (having lost weight)	riutunut
rabble	lauma (herd, flock)	roskaväki
corroborate strengthen	vahvistaa	-entry removed- -entry kept-
cede relinquish	luovuttaa (to give up)	-entry kept- -entry removed-
rape	raiskaus (N) raiskata (V)	-entry kept- -entry added-

Table 5.2: Examples of fixes to the SELF and FEIL lexicons

In some cases the word had been clearly mistranslated or the connotations were wrong. A common case was where the original English word was more colloquial in nature and context was not available to aid the translation. A typical example were the negative connotations of the word *birch* which had been translated as *koivu* in Finnish. *Birch* in this context was clearly taken to mean “a birch rod or bundle of twigs for flogging” or the verb associated with using a birch to flog, i.e. to birch, and not the tree.

This was reflected in the intensity scores, but not the translation.

For *emaciated*, the automatic translation was too general. The corrected form is less common but almost identical in meaning and connotation to the original English. As the term *rabble* has such negative connotations in English, the automatic translation was much too neutral and was changed to a word with similar connotations. As for *corroborate* the issue was with there not existing a one-word translation for *corroborate* in Finnish and the meaning of the automatically translated word being much closer to *strengthen*. The meaning of the original English words *cede, relinquish* are somewhat synonymous so it makes sense to have both words in the English lexicon with nearly identical emotional intensities, but there is only one one-word translation for Finnish so the duplicate was removed.

In some cases the Finnish translation had been inflected and needed to be changed to the word's basic form. A common issue was the English word being polysemic or homonymous and the Finnish translation different for these words. In some cases, where the intended meaning of the word was obvious from the intensity scores, we corrected the translations, in others we made two entries for Finnish for both meanings and used the intensity score of a synonymous word if the intensity scores seemed plausible.

It is a slippery slope to try and push one's own judgment onto the intensity scores and emotion-word associations, therefore, I avoided making any judgments of my own. In some cases this was not possible, so I carefully chose a near-synonym and used its intensity score and word-associations where applicable. I rather deleted a word than gave it a subjective score based on my own assumptions and cultural biases.

On the topic of cultural biases, I found another type of mistranslation: those where the cultural associations of words did not match. The original data was annotated by mainly North-Americans and therefore the scores and associations align with North-American culture. The discrepancy was most noticeable with religious words (see table 5.1 for a related example), as Finland is notably less religious than the US and Canada (Skirbekk et al., 2015).

The final distribution of the SELF lexicon can be seen in table 5.3. All the adjustments mentioned in this section resulted in an overall reduction in lexicon size by 12.2% from 14182 word-emotion association pairs to 12448 entries in Finnish.

positive	negative	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
2117	2938	1084	783	919	1309	636	1059	473	1130

Table 5.3: Distribution of Emotions in SELF: Sentiment and Emotion Lexicon for Finnish

The final distribution of the FEIL lexicon can be seen in table 5.4. All the adjustments mentioned in this section resulted in an overall reduction in lexicon size by 10.5% from 8149 intensities to 7291 entries in Finnish.

anger	anticipation	disgust	fear	joy	sadness	trust
1304	805	946	1554	1145	183	1354

Table 5.4: Distribution of Emotions in FEIL: Finnish Emotion Intensity Lexicon

The SELF and FEIL lexicons were used in papers IX and X.

A good emotion lexicon can improve the performance of a rule- or lexicon-based emotion detection system, but can also be used in conjunction with machine learning approaches to create hybrid models. Lexicon-based methods are especially good for tasks that require a highly-qualitative analysis of the data. For more purely quantitative tasks, however, lexicon-based methods have shortcomings as can be seen in paper V, for example. Therefore, quality emotion annotated datasets for languages other than English are needed in order to evaluate data-driven approaches to emotion detection.

5.2 Fine-grained Emotion Annotation

Annotated datasets are required for supervised machine learning and the evaluation of models. There are very few emotion annotated datasets and the few that do exist are almost all for the English language. This was even more true when I started my doctoral research, but it holds true even today, particularly for languages other than English.

In order to annotate such datasets, a few things are required:

(i) Source data: something to annotate

(ii) Annotation scheme: how to annotate

(iii) Annotators: someone to do the annotations

(iv) Annotation platform: an annotation tool

From the beginning I knew I wanted to create resources for languages other than English and Finnish as well, so for the data (i) I chose the OPUS parallel corpus as it would allow me to later compare the annotation between different languages. For annotation scheme (ii) I chose Plutchik's core emotions as these were commonly used in other studies and therefore would make comparison easy, but also because I was already familiar with the underlying theory from paper V. I have now taught the course *Introduction to Language Technology* seven times (see Öhman 2019), and knew that the students could learn more about sentiment analysis, annotation, and supervised machine learning if they could partake in a practical annotation task. The students became my annotators, and the annotator experience is discussed in detail in paper IV. No suitable platform existed, so I set out to design such a platform. With the help of my graduate student, Kaisla Kajava, we created *Sentimentator*, which will be discussed in this section.

Sentimentator is a light-weight dockerized open source[1] web-platform[2] created with Python flask (Grinberg, 2018). Annotations are stored in an SQLite database together with relevant metadata. The *Sentimentator* annotation platform was first introduced in paper II as a streamlined multi-dimensional annotation platform for sentiment and emotion annotation of individual lines of movie subtitles. In paper III we propose further improvements to *Sentimentator*, specifically a self-perpetuating algorithm based on the annotation of seed sentences and an inverted Plutchik's wheel for annotation (figure 5.2).

We suggest that by annotating by clicking on Plutchik's wheel directly, it is possible to get more accurate annotations with intensities included as the annotators might feel less conflicted about judgments when they can adjust their score to reflect, e.g. *mild remorse*, by clicking somewhere between *pensiveness* and *boredom*. A score that previously might have been marked as *skip*, *neutral*, or either *sadness* or *disgust* or even both, now more accurately reflects the true nature of the sentence.

The diagram in figure 5.3 describes the annotation process with *Sentimentator*. First expert annotators annotate seed sentences that act as a gold standard for annotations.

[1] `https://github.com/Helsinki-NLP/sentimentator`

[2] `http://sentimentator.it.helsinki.fi`

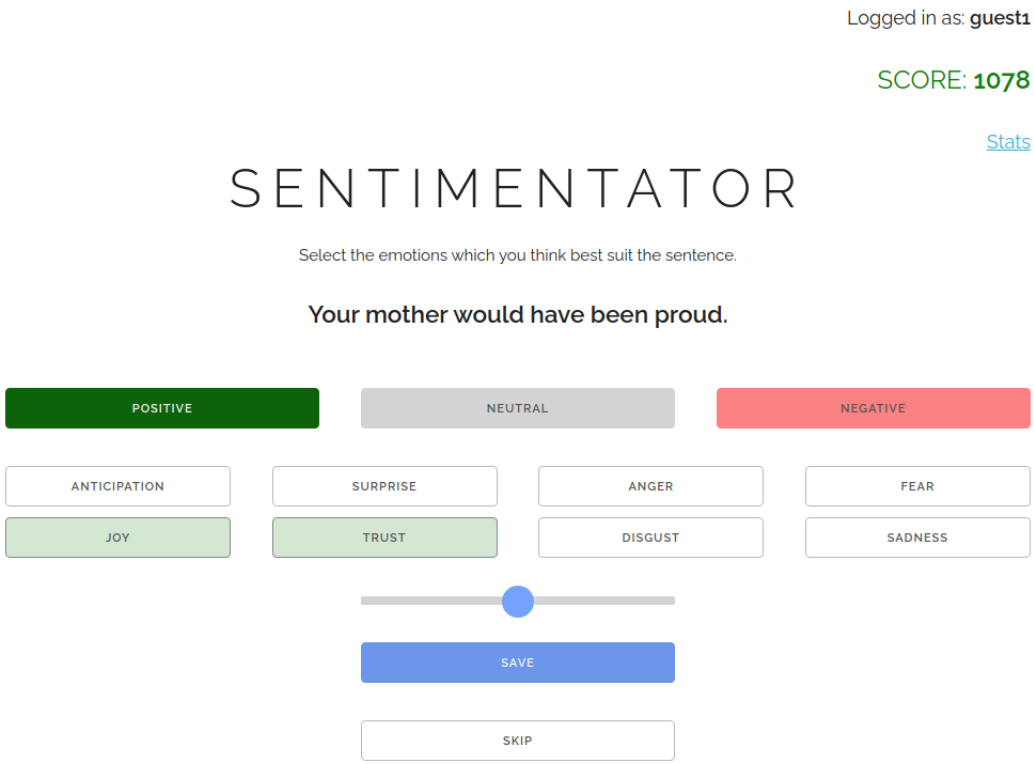

Figure 5.1: Screenshot of Sentimentator's interface

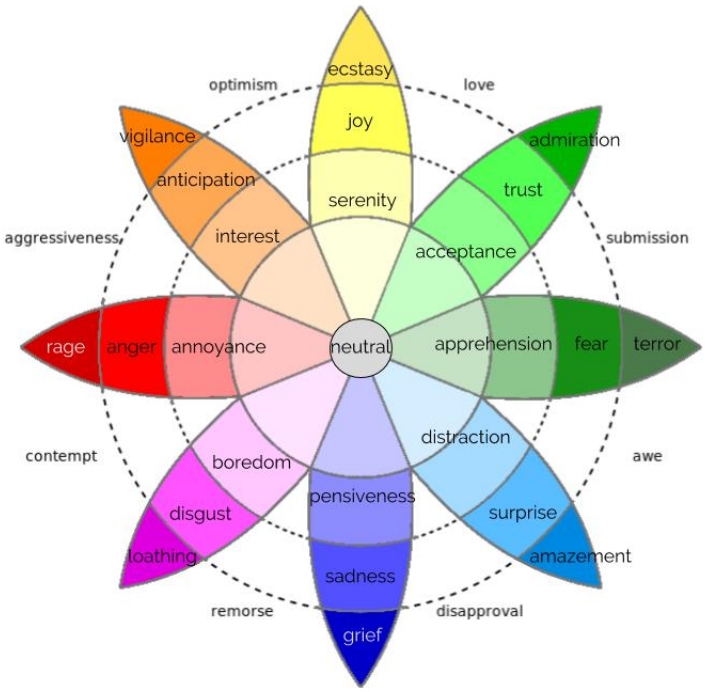

Figure 5.2: Inverted Plutchik's wheel

Then crowd annotators annotate sentences, both seed sentences and non-seed sentences. The crowd annotators' rank increases as their annotation of seed sentences matches the expert annotation of the seed sentence. Once a crowd annotator has high enough rank, their annotations are considered pseudo-seed sentences producing a confidence score against which other crowd annotators are compared.

Players accumulate rank (R), which is a prestige or reliability score based on how well the player's annotations correlate with seed sentences and pseudo-seed sentences.

$$R = \frac{T_{sv}}{V_{max}}$$

where T_{sv} stands for total score from validated sentences and V_{max} for maximum possible score for the player in question from validated sentences.

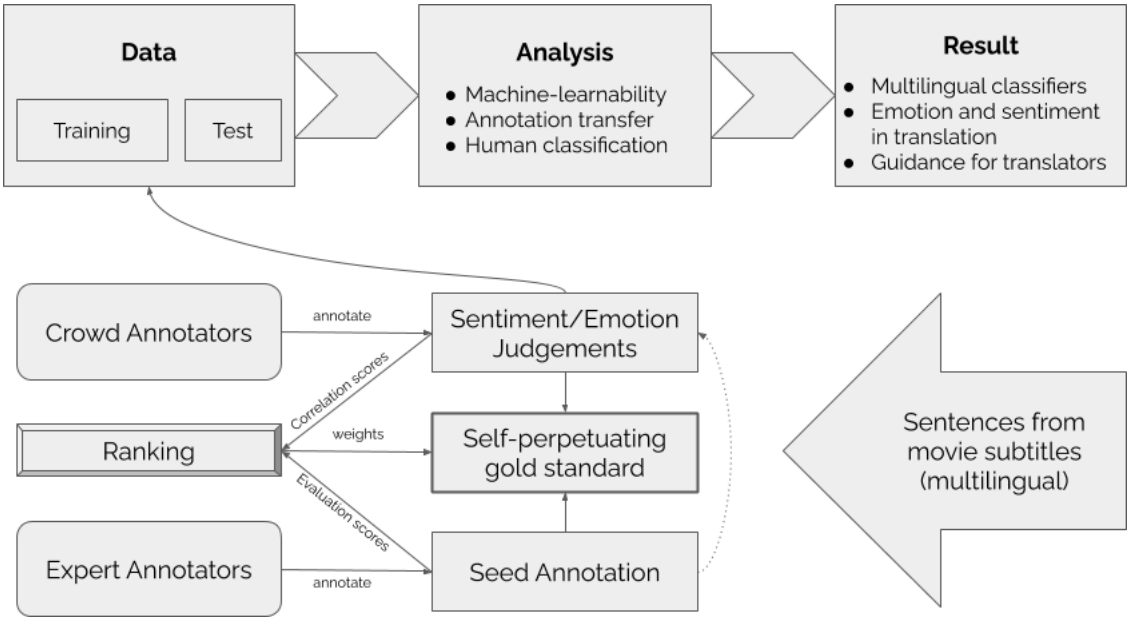

Figure 5.3: Diagram of *Sentimentator* system design

The relationship between the score based on peer annotation (S_p) and the ranking of the annotator (R_{oa}) who has annotated the same sentence before influences the score for the annotation as follows:

$$S_p = \frac{P_s}{S_{oa}} \times R_{oa}$$

where P_s stands for the pre-adjustment annotation score of the peer and S_{oa} stands for the score of the other annotator. In practice this will work much like a weighted average across all peers. The number of annotations per sentence is also limited (see paper II and VI).

In paper IV annotators indicated that they found the platform clear and easy to use, however, the task remained monotonous for many. Further gamification and in-game community building would likely ease the feeling of repetitiveness for the annotators. Nonetheless, many annotators expressed that the task helped them understand machine learning and the nature of emotions in text better and that the task became more interesting the more they annotated.

Perhaps the challenge is not all about making the annotations “fun”, although that

certainly helps, but to choose annotators based on some kind of initial screening process that would attempt to quantify the aspects of personality that make someone more prone to find enjoyment in the process of learning.

5.3 Preservation of Sentiments and Emotions in Translation

Because I want to support as many languages as possible, but do not have the resources to create annotations for all of them, I wanted to find a way to bootstrap more annotations and tools for languages other than English. This meant that I needed to study transfer learning and annotation projection. However, before doing that, I needed to find out how well emotions are preserved in translation and therefore, whether transfer learning was possible for emotion detection tasks. In this section I present my findings about sentiment and emotion preservation.

In papers V and VI sentiment and emotion preservation in translation is examined. In paper V this is done using lexicon-based methods and calculating cosine similarity for language vectors; I wrote a script that would create a multidimensional array of texts from which emotions could be extracted into ten-dimensional vectors (see table 5.4). This script uses a simple approach where I determine the Emotions E of each input unit U by calculating the sum of each emotion e from cross-referencing an emotion or emotion intensity lexicon L with the lemmas in U. This sum Σ is then normalized by the word-count W in U. The process is identical for emotions and intensities except that for emotions the value is always either 0 or 1 and for intensities the value is between 0 and 1. These results are then vectorized, meaning that each dimension of the vector corresponds to a specific emotion, so that cosine similarity measures can be obtained to compare the similarity of documents. These vectors can then be summed to create an overview of emotion-word content and distribution normalized by word count in the document. So in essence this is word-level granularity extrapolated to document-level and in some cases language-level (i.e. all the documents in a specific language). The script is in essence simply matching words from the documents to words in the emotion lexicon. This process is outlined in figure 5.4.

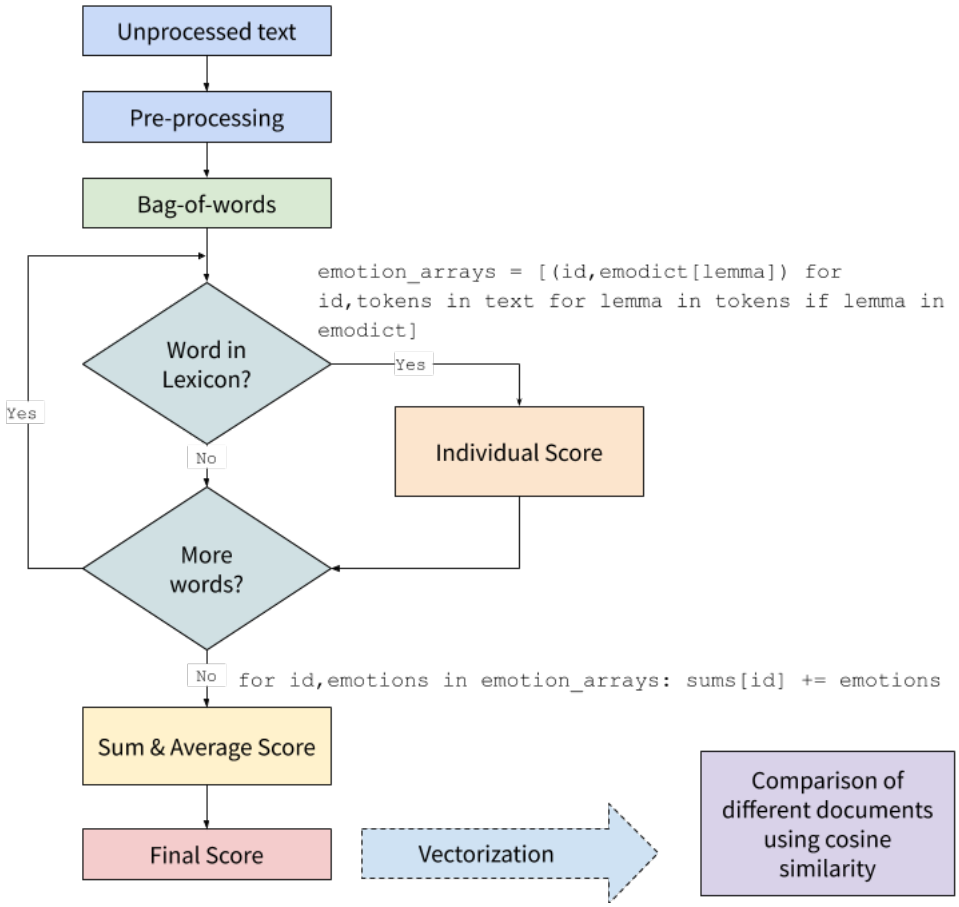

Figure 5.4: Lexicon-based methodology[3]

The classification models used in papers III and VI relied on Multilayer Perceptron and Multinomial Naïve Bayes. We selected Multinomial Naïve Bayes (NB), Multilayer Perceptron (MLP), SVMs, and Maximum Entropy classifiers for our experiments. The data was split into 90:10 stratified training and test sets. The classification results can be seen in table 5.3.

The input data was lemmatized using UDPipe (Straka et al., 2016), or similar models optimized for specific languages where available, on the Center for Scientific Computing[4] (CSC) supercomputer *Taito*. The lemmas were then extracted and used as the input

[3]Figure loosely inspired by Kolchyna et al. (2015) fig. 1.

[4]`https://www.csc.fi/`

	Classifier	Accuracy	
		Binary	*Multi-class*
EN	MLP	0.7090	0.4931
	MNB	0.7152	0.5084
	LinearSVC	0.7213	0.5100
	MaxEnt	0.7198	0.5008
FI	MLP	0.6493	0.3767
	MNB	0.6646	0.4089
	LinearSVC	0.6417	0.3890
	MaxEnt	0.6462	0.4058
FR	MLP	0.6493	0.3767
	MNB	0.6692	0.3997
	LinearSVC	0.7029	0.4089
	MaxEnt	0.6876	0.3950
IT	MLP	0.6692	0.4150
	MNB	0.6784	0.4272
	LinearSVC	0.6815	0.4288
	MaxEnt	0.6953	0.4166

Table 5.5: Overall classification accuracy on the testing set.

data. Once these words were vectorized and summed per language, I calculated some similarity measures, mainly cosine similarities between the vectors. Cosine similarity is a measure of how similar the direction two vectors are pointing at is, with 1 being the same direction, 0 perpendicular, and -1 the opposite direction. That means that the closer the value is to 1, the more similar the vectors. This approach was used in paper V, where I examined sentiment and emotion preservation across languages. Table 5.6 shows how the different emotions in the different language pairs retained sentiment and emotion information.

	Emotion / Sentiment										
	pos.:	neg.:	anger:	anticip.:	disg.:	fear:	joy:	sad.:	surpr.:	trust:	ALL
Language	Movie Subtitles										
EN-FI	.6051	.4535	.7507	.8299	.7761	.7818	.8364	.7766	.8964	.7471	$0.752^{\pm 0.232}$
EN-SV	.5709	.4744	.7897	.8310	.7817	.7948	.7710	.7865	.8922	.7631	$0.802^{\pm 0.220}$
ES-PT	.6186	.4912	.7964	.8419	.8119	.7715	.8749	.7299	.9248	.8251	$0.746^{\pm 0.231}$
Language	EuroParl										
EN-FI	.5670	.4613	.7733	.8138	.7839	.7805	.8240	.7755	.8914	.7434	$0.788^{\pm 0.241}$
EN-SV	.3219	.4028	.7148	.6590	.7420	.6605	.7314	.6902	.7888	.4851	$0.665^{\pm 0.213}$
ES-PT	.4172	.4480	.6849	.6934	.7815	.6783	.7352	.6501	.8278	.5570	$0.692^{\pm 0.178}$
AVG:	.5168	.4552	.7516	.7782	.7795	.7446	.7955	.7348	.8702	.6868	

Table 5.6: Accuracy of matched sentiments and emotions across languages according to lexicon-based classification. *ALL* refers to the averaged cosine similarity of the 10-dimensional sentiment/emotion vectors and the number in superscript gives the standard deviation observed in the data. The abbreviations for the languages are by ISO code: EN - English, ES - Spanish, FI - Finnish, PT - Portuguese, and SV - Swedish

The results indicate that for most parts and most language pairs, the OPUS subtitle data preserves sentiments and emotions better than the EuroParl datat. Perhaps the modality of the data lends itself to better retained expressed emotions.

The most common emotions in the texts were the same for all languages: *negative*, *positive*, then fairly similar for *trust*, *disgust*, *anger*, *fear*, *joy*, *sadness*, and generally much lower for *anticipation* and *surprise*. This is most likely related to the higher cross-language agreement for these emotions: the more common a sentiment/emotion, the more chances of one language detecting it but it being missed by the other and therefore

decreasing the cross-language agreement score.

Finally, cosine similarity and Euclidean distance measures were calculated for the two language vectors. For all measures mean, median, and standard deviation are presented. Table 5.7 below shows these similarity scores.

	EN-FI			EN-SV			ES-PT		
Cosine	0.746	0.775	0.231	0.752	0.795	0.232	0.802	0.868	0.220
Euclidean	1.458	1.414	1.153	1.400	1.414	1.157	1.199	1	1.097
	mean	median	std.dev.	mean	median	std.dev.	mean	median	std.dev.

Table 5.7: Similarity scores and distances

The cosine similarity scores indicate that the Finnish and English vectors are the most dissimilar, with only slightly higher similarity scores for the English-Swedish pair. The Spanish-Portuguese scores, however, show higher similarity scores than either of the other two languages. This might very well be because of either the similarities of the two languages, or even due to the cultural similarities of emotion expression.

In paper VI we explored how emotions are preserved in translation using both human translated pairs and automatic classification. We found that with Finnish as the target language, emotions were best preserved in terms of accuracy (0.88-0.90, table 5.8) and inter-annotator agreement was also higher on average (see table 5.9). French accuracy scores were a few percentage points lower, and Italian scores a little lower yet, so linguistic similarity is an unlikely explanation here.

Nonetheless, sentiments and emotions were preserved quite well in translation. For English and Finnish, *joy* was the emotion that was best preserved, and for Italian *sadness* was best preserved. Despite the fact that machine learning approaches have found that *anger* is typically the easiest to predict (see chapters 2 and 3), in this case it was not preserved that well. *Surprise*, on the other hand, is typically the most difficult emotion for classifiers to predict correctly, and in this case too it was the least preserved. Here too annotators most often found it difficult to distinguish between *anger* and *disgust*.

The Cohen Kappa scores displayed in table 5.9 were at a moderate level of agreement (between 0.40-0.59) for all sentiment and emotion classes, which might even be considered good for such a complex and subjective task of annotation work.

	pos	***neg***	*ang*	*ant*	*dis*	*fea*	*joy*	*sad*	*sur*	*tru*	***Total***
EN → FI	0.87	0.86	0.81	0.87	0.88	0.91	**0.92**	0.87	0.74	0.86	***0.86***
EN → FR	0.83	0.83	0.78	0.83	0.88	0.83	**0.90**	0.87	0.73	0.82	***0.83***
EN → IT	0.81	0.83	0.81	0.76	0.83	0.84	0.86	**0.92**	0.70	0.80	***0.82***

Table 5.8: Sentiment and emotion preservation accuracy.

	ang	*ant*	*dis*	*fea*	*joy*	*sad*	*sur*	*tru*
EN → FI	0.45	0.47	0.47	0.48	0.48	0.46	0.44	0.46
EN → FR	0.44	0.46	0.47	0.45	0.48	0.46	0.43	0.45
EN → IT	0.44	0.43	0.45	0.42	0.43	0.48	0.40	0.43

Table 5.9: Inter-annotator agreement per language pair (Kappa).

ang	*ant*	*dis*	*fea*	*joy*	*sad*	*sur*	*tru*
0.82	0.84	0.81	0.82	0.81	0.83	0.80	0.81

Table 5.10: Inter-annotator agreement in the English data set.

As presented in table 5.10, each sentiment and emotion class in the English base data set had a Cohen Kappa score of strong agreement. This indicates an annotation process reliable for evaluating sentiment preservation for the purposes of this study.

5.4 Annotation Projection and Transfer Learning

As I have now shown that emotions are preserved sufficiently in translation, I can go ahead and test annotation projection. I used the student annotated dataset discussed earlier in section 5.2 to project annotation from English onto 43 other languages. I call this dataset XED[5] (for CROSSlingual Emotion Detection). I evaluate the English dataset thoroughly as I needed to first make sure that the source data was of sufficiently high quality before the annotations could be projected onto other languages. I will first discuss the English dataset and then evaluate the projections.

For XED we used BERT (Devlin et al., 2019) to evaluate the classifiability of our

[5] https://github.com/Helsinki-NLP/XED

annotations. BERT has in recent SemEval tasks shown to be the best performer in many different types of tasks as it has consistently outperformed other models recently (see e.g. section 4.3, Zampieri et al. (2020), Yang et al. (2020), and Patwa et al. (2020)). We use a stratified split of 70:20:10 for training, dev, and test data, and for some datasets also experimented with scikit-learn's (Pedregosa et al., 2011) built-in stratified split function.

For English we used uncased BERT fine-tuned for the classification task with a batch size of 96. The learning rate of Adam optimizer was set to 2e-5 and the model was trained for 3 epochs. The sequence length was set to 48. We perform a 5-fold cross validation. For Finnish we use FinBERT cased (Virtanen et al., 2019) with the same parameters as with English. We showed in paper VIII that language-specific BERT models tend to outperform multilingual BERT, and the same is true for Finnish. The developers show an increase in performance of 2.11 percentage points from 90.29% accuracy with multilingual BERT to 92.40% with Finnish BERT (Ruokolainen et al., 2019) on document classification tasks.

The XED dataset (paper VII and partly III; IV; VI) was collected in three different rounds using a combination of trained expert annotators and student annotators annotating on the *Sentimentator* platform (papers II; III). They annotated context-free lines from OPUS (Lison et al., 2019). The original expert annotations were for English, French, Italian, and Finnish. The student annotations were mostly for English, but with a significant amount of Finnish annotations, and a very small minority of Swedish, Russian, Italian, and French annotations.

The annotations were made using Plutchik's basic emotions as a base resulting in 8 distinct emotion categories plus *neutral*. The granularity level is not exactly sentence-level as the source data is movie subtitle lines which have been pre-processed to merge sentence fragments that span more than one line with automatic sentence boundary detection when such appear in the middle of a subtitle frame. One reason for the short and incomplete sentences is perhaps the fragmented nature of human language, another the artistic choices of the creators of the original manuscript and of the subtitles. Some of the shortest "sentences" are only one character long (e.g. *!*) and the longest can consist of up to three sentences.

The annotations were extracted and I removed both noisy annotations and noise in

the data itself; we wanted 3 annotators to agree on an emotion label, but as annotators were allowed to skip sentences this was not always the case. For the final dataset, sentences were included if at least 2 annotators agreed on the label. A few instances that were annotated by a single annotator, but checked by expert annotators during the preprocessing stage, were included. Additionally, those sentences annotated as both *neutral* and an emotion were removed from the *neutral* set. The noise removed from the text itself was mostly superfluous characters such as quotation marks.

I also ran Stanford NER (Finkel et al., 2005) on the data to replace locations and people, but not institutions, as I figured the emotions associated with a line like "He received a letter from the IRS" are very different to "He received a letter from Harvard", but a sentence like "Sara always wanted children" is basically identical to "Christa always wanted children"[6], especially since the data was annotated out of context.

Once the data was clean and the annotations were cleaned up, I used the GoEmotions (Demszky et al., 2020) template for input data:

I am angry	1
I am looking forward to Friday's party!	2,5
There's a spider in your hair!!!	3,4,7

The numbers stand for Plutchik's basic emotions in alphabetical order. We reserved 0 for *neutral*, so 1 is *anger*, 2 is *anticipation* etc.

The annotated XED dataset for English contains 24164[8] annotations[9], excluding *neutral*, of 17530 unique sentences. It is one of the largest emotion datasets for English that

[6]Although Kiritchenko and Mohammad (2018) found that many sentiment and emotion analysis systems favor certain types of names, this was related to sexism and racism.

[7]By *active* I mean annotators who annotated over 100 lines.

[8]Note that the total number of annotations excluding *neutral* (24164) and the combined number of annotations (22424) differ because once the dataset was saved as a Python dictionary, identical lines were merged as one (i.e. some common movie lines like "All right then!" and "I love you" appeared multiple times from different sources).

[9]By this I mean that a sentence could have been annotated as containing 3 different emotions by the annotator. This would count as 3 annotations on one unique data point.

Number of annotations:	24164 + 9384 neutral
Number of unique data points:	17530 + 6420 neutral
Number of emotions:	8 (+pos, neg, neu)
Number of annotators:	108 (63 active[7])
Number of labels per data point:	1: 78% 2: 17% 3: 4% 4+: 0.8%

Table 5.11: Overview of the XED English dataset

we are aware of[10]. A majority of the sentences for English were assigned one emotion label (78%), 17% were assigned two, and roughly 5% had three or more categories (see also table 5.11).

The final dataset contained 17530 unique emotion-annotated sentences as shown in table 5.11. There are a further 6420 sentences annotated as *neutral* making the full size of the English dataset 23950 unique annotated sentences[11]. The distribution of the emotion labels are shown in table 5.12.

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	Total annotations
4182	3660	2442	2585	3139	2635	2635	2886	24164
17.31%	15.15%	10.11%	10.70%	12.99%	10.90%	10.90%	11.94%	XED percentage
15.09%	10.15%	12.80%	17.86%	8.34%	14.41%	6.46%	14.89%	the NRC Emotion Lexicon percentage

Table 5.12: Emotion label distribution in the XED English dataset

If we compare the number of annotations per emotion, the distribution is fairly balanced. Only *anger* and *anticipation* are slightly over-represented. If we compare the distribution of XED to the NRC Emotion Lexicon (Mohammad and Turney, 2013) there are a few differences. Especially *fear*, which is the largest category in the NRC lexicon but the second smallest in XED, and *joy* which takes up almost 13% of XED annotations but only 8% of NRC. I would suggest that this is due to the differences in source

[10]GoEmotions (Demszky et al., 2020), CrowdFlower, and TEC (Mohammad, 2012) are bigger when excluding *neutral* for XED. Including *neutral*, XED is on par with these larger datasets.

[11]This is after some further clean-up of duplicates etc.

data. The NRC lexicon is built using mainly dictionaries, whereas movie subtitles are probably more diverse due to the many different genres represented. Although *anticipation* and *trust* are considered positive emotions, *joy* can also be seen to be the most objectively *happy* category and therefore likely to be chosen as a label when an annotator is determining on the most suitable label for anything remotely positive that can not immediately be associated with the other positive categories.

As most of our data was annotated for emotions only[12], we needed to map these emotions to sentiments. In our scheme and with the annotation guidelines in mind, we mapped *anger*, *disgust*, *fear*, and *sadness* as *negative*, and *anticipation*, *joy*, and *trust* as *positive*. *Binary* refers to *positive* and *negative*. *Surprise* was either discarded or included as a separate category[13] (see table 5.13), because of how it is expressed in text (see chapter 2 and 3).

Evaluation Metrics

We achieve macro f1 scores of 0.536 for our multilabel classification with a fine-tuned BERT model. Using named-entity recognition increases the accuracy slightly from 0.530 to 0.536. For binary data mapping emotions to the sentiments of positive and negative (non-multiclass, non-multilabel classification) our model achieves a macro f1 score of 0.838 and accuracy of 0.840.

Looking at table 5.13 it is clear that the more categories or classes the dataset consists of and the model has to predict, the lower the accuracy and macro f1 score, suggesting that the model has trouble distinguishing between certain classes. The results also indicate that we were justified in using NER with our data as it improves accuracy by 0.6 percentage points. Therefore, all the other classification experiments used the data where names and locations had been replaced by generic tags, including for the other languages.

The inclusion of *neutral* is tricky at best. As mentioned in chapter 2, previous re-

[12]We do have a few sentiment-only annotations as the platform does not force you to choose an emotion after selecting a sentiment, but they are a very small minority.

[13]Some papers have used *positive surprise* and *negative surprise* separately, e.g. Alm et al. (2005), whereas others have ignored *surprise* completely.

data	f1	accuracy
English without NER, BERT	0.530	0.538
English with NER, BERT	0.536	0.544
English NER with neutral, BERT	0.467	0.529
English NER binary with surprise, BERT	0.679	0.765
English NER true binary, BERT	0.838	0.840
English NER, one-vs-rest Linear SVC	0.502	0.650-0.789 / class

Table 5.13: Evaluation results of the XED English dataset

search has shown that neutral sentences increase accuracies, however, looking at our results, it seems that neutral sentences should perhaps be ranked somehow from most to least *neutral* and n of the most neutral sentences should be used, where n is the average number of sentences marked as belonging to the other classes. This would balance the classes better. As it currently stands, *neutral* is such a large class that simply categorizing everything as *neutral* would yield accuracy scores of around 0.27-0.34 depending on the dataset. Comparing *English with NER* and *English NER with neutral* suggests that in our dataset the inclusion of *neutral* only confuses the learning/training algorithm. This is likely due to both the data imbalance, but even more so the nature of the data. In paper IV many annotators commented on how many sentences felt like they could be either *neutral* or a low intensity of one of the eight emotions.

The confusion matrix (see Figure 5.5) of the single[14] label annotations, reveals that *disgust* is often confused with *anger*. *Disgust* is in fact more often classified as *anger* than it is *disgust*. Although *anger* is also most often misclassified as *disgust*, the ratio of misclassifications is very different. *Disgust* is indeed the hardest emotion to classify correctly. *Joy*, *anger* and *anticipation* are the labels that are classified correctly the most if going by the confusion matrix, this might simply be a reflection of their prevalence in the training data as the dataset is not fully balanced.

I wanted to examine the different emotions individually so I calculated precision and recall for each emotion, and as speculated in section 5.4, *joy* has both the highest

[14]We use the expert annotated seed dataset for this as a confusion matrix for multilabel classification would defeat the point.

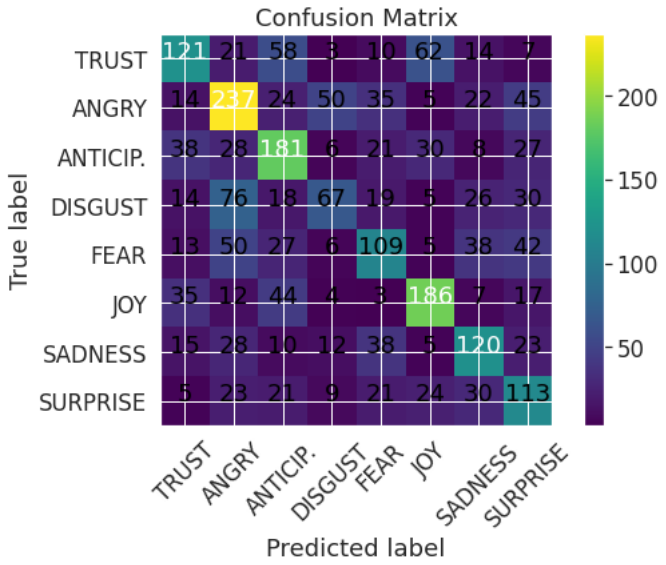

Figure 5.5: Confusion matrix for single-label part of the XED English dataset

precision and recall. For future experiments it might be interesting to test different combinations of merging *anger* with *disgust* and/or possibly with *fear* and *surprise* as well. I say this because *anger* is the largest class in the dataset, yet it is only second in precision, and third in recall suggesting that it is very much confused with *disgust*, *fear*, and *surprise* (cf. figure 5.5). Interestingly, despite the low precision for *surprise*, recall is relatively high at 0.459 (*disgust* has the worst recall rates).

	anger	anticip.	disgust	fear	joy	sadness	surprise	trust
precision	0.499	0.473	0.427	0.426	0.578	0.453	0.372	0.475
recall	0.549	0.557	0.263	0.376	0.604	0.478	0.459	0.409

Table 5.14: Precision and recall for each emotion in XED

5.4.1 Multilingual Approaches

From our source data we can extract parallel sentences for 42 languages and language variants[15]. For 11 of these languages there are more than 10,000 sentences available for projection as per table 5.15, these languages are Italian, Finnish, French, Czech, Portuguese, Polish, Serbian, Turkish, Greek, Romanian, Spanish, and Brazilian Portuguese.

it	fi	fr	cs	pt	pl	sr	tr	el	ro	es	pt_br
10582	11128	11503	11885	12559	12836	14831	15712	15713	16217	16608	22194

Table 5.15: Languages with over 10k parallel sentences with our annotated English data

We test our annotation approach first and foremost on Finnish as we have some annotated Finnish data and can therefore directly compare the different approaches. This means that we can look at the differences between (1) manually annotated data between different languages (English and Finnish), (2) projected annotations and manual annotations for Finnish, and (3) machine translated English into Finnish compared with English, Finnish projections, and machine translated Finnish.

The Finnish data consists of 19,529 individual annotations and 14,449 unique emotion annotated sentences plus an additional 7536 sentences annotated as *neutral*. 77% of the sentences had only one label, 18% two labels, 5% three labels, and only 1.2% had four or more (see table 5.16 for more details). The projected dataset has 11,128 data points.

We have made all 43 datasets available on GitHub[16], but only evaluate Finnish as proof of concept.

Comparing the distribution of the emotion labels in the Finnish and English dataset shows that labels have a balanced distribution in the Finnish dataset too, although *anticipation* if slightly less common in the Finnish than in the English data.

[15]The full list of available languages: Arabic, Bulgarian, Bosnian, Czech, Danish, German, Greek, Spanish, Estonian, Basque, Persian, Finnish, French, Galician, Hebrew, Croatian, Hungarian, Indonesian, Icelandic, Italian, Japanese, Macedonian, Malayalam, Malay, Dutch, Norwegian, Polish, Portuguese, Brazilian Portuguese, Romanian, Russian, Slovak, Slovenian, Serbian, Swedish, Thai, Turkish, Vietnamese, Mandarin Chinese (Simplified), Taiwanese Mandarin Chinese (Traditional)

[16]`https://github.com/Helsinki-NLP/XED`

Number of annotations:	21984 (/w neutral)
Number of unique data points:	14449 + 7536 neutral
Number of emotions:	8 (+neu)
Number of annotators:	40 active
Number of labels per data point:	1: 77% 2: 18% 3: 5% 4+: 1.2%

Table 5.16: Overview of the XED Finnish dataset

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	
3345	2496	2373	2186	2559	2184	1982	2404	19529
17.13%	12.78%	12.15%	11.19%	13.10%	11.18%	10.15%	12.31%	

Table 5.17: Emotion label distribution in the XED Finnish dataset

For the annotation projection data I used the 11,128 Finnish sentences that were directly parallel to the annotated English dataset, and for the machine translated dataset I used Google Translate[17] to automatically translate these sentences in order to be able to compare these directly with the results from the projected annotations. The model was trained with these datasets and evaluated on the manually annotated dataset.

data	f1	accuracy
Finnish annotated	0.5070	0.5131
Finnish annotated, coarse	0.6135	0.6891
Finnish projected	0.4461	0.4542
Finnish MT	0.4481	0.4556

Table 5.18: Annotation projection evaluation results of the XED Finnish dataset using FinBERT

The annotated dataset fared the best; we used FinBERT cased to gain accuracies of 0.51 for 8 labels for the manually annotated data. I also did a preliminary test with

[17] Accessed October 1, 2020.

combining some of the categories, namely *anger, disgust, fear* as one class, and the removal of *surprise*, resulting in 5 classes and an accuracy of 0.69 (see results for *Finnish annotated, coarse*). This demonstrates that *anger, disgust,* and *fear* are similar enough to allow such a combination approach if suitable for a project. For Finnish MT I used the English manual annotation data and used Google Translate to acquire an comparable Finnish dataset. For both the projected and machine translated Finnish datasets, the same test set, namely a subset of the manually annotated Finnish set where overlap with the projected and machine translated datasets had been removed, was used for optimal comparability.

Table 5.19 shows projection results for a selection of language where evaluations were conducted using language-specific BERT models available from Huggingface transformers.

data	f1	accuracy
Finnish projected	0.4461	0.4542
Turkish projected	0.4685	0.5257
Arabic projected	0.4627	0.5339
German projected	0.5084	0.5737
Dutch projected	0.5155	05822
Chinese projected	0.4729	0.5247

Table 5.19: Annotation projection evaluation results for other languages

The accuracies of the projected datasets and the machine translated dataset were not drastically worse than accuracies for the manually annotated one, in line with previous research. The accuracy and f1 scores for the other projected languages are also very similar to the Finnish projections, suggesting that the projections are of similar quality in all the tested languages, and by extension likely in the other languages as well, although there may be differences and exceptions due to e.g. the nature of the original translations. Unlike the Finnish dataset, which was not cross-validated because it was tested on the same expert annotated dataset as the machine translated dataset, for the other non-English datasets I used 5-fold cross-validation to offset the smaller sizes of the datasets.

Interestingly, the German and Dutch datasets achieve higher accuracy scores than even the original annotated English. There seems to be no clear reason for this, and the German dataset was only about a quarter of the English dataset, so it would have been more likely to see much lower accuracies. All the datasets were evaluated in an identical manner. The only difference being the language-specific BERT models, so it is unlikely the scores reflect a bug in the code. It is much more likely that the high accuracy is a reflection of the dataset balance and size, and that the above-average f1 scores are a reflection of the typological, cultural, and linguistic similarities of the three Germanic languages of English, German, and Dutch resulting in the original translations being more similar. When the translations are similar, it is likely that less information loss occurs and therefore the projected annotations better describe the emotional content of the subtitle.

These results suggest that both machine translation and annotation projection are viable approaches when manually annotated data is not available, and that using machine translation can be a viable approach to increase the number of data points in small or imbalanced datasets where data is available in another language.

5.5 Fine-grained Emotion Annotation and Practical Applications

After several papers dealing with sentiment analysis from a research point of view, I wanted to see how the different sentiment analysis and emotion detection methods I had created fared when applied to real-world tasks. That is I felt that it was important to show the usefulness of sentiment analysis and emotion detection outside of pure computer science or computational linguistics. The downstream applications here range from theoretical offensive speech detection for SemEval, to literature analysis and political rhetoric.

5.5.1 OffensEval

A first trial came with the SemEval 2020 task of offensive speech detection (task 12, OffensEval). The main objective was to use the OLID dataset (Zampieri et al., 2019b) to complete the subtasks to the highest possible accuracy (macro f1 score). The subtasks were:

A. Offensive Language Identification for Arabic, Danish, English, Greek, Turkish: whether a tweet is offensive or not.

B. Categorization of Offense Types: whether an offensive tweet is targeted or untargeted.

C. Offense Target Identification: whether a targeted offensive tweet is directed towards an individual, a group, or other.

Our team placed first in subtask A, where our focus was Danish. The task was to correctly predict whether a message was offensive (OFF) or not (NOT). We compared Nordic BERT with the commonly used multilingual BERT, and found that at least in this case the language-specific model performed better. We also tested a RandomForest classifier with TFIDF due to the size of the dataset. For random forest we did a lot of pre-processing and used a 10:1 split with 10-fold cross-validation to improve the reliability of our results and used grid search to find the optimal parameters. For BERT we compared different pre-trained models and parameters to find the optimal solution. We found that Nordic BERT[18] outperformed multilingual BERT, indicating that further pre-training can have a positive boost on BERT's performance. The results with Nordic BERT were clearly the best and were thus used to achieve our final submission score (see table 5.20).

Danish was probably the most difficult of the target languages in this subtask due to the dataset being the smallest. Once we realized that the number of emotion words in offensive posts were significantly higher than in non-offensive posts, even for *joy* (see table 5.21), we attempted to incorporate emotion weights into the data by using parts of the Google Translated NRC Emotion Lexicon (Mohammad and Turney, 2013) as

[18] `https://github.com/botxo/nordic_bert`

System	**macro-F1**
All NOT	0.465
Random Forest with TFIDF	0.773
Multilingual BERT-Base	0.768
Nordic BERT	0.804

Table 5.20: Development results sub-task A, Danish dataset

features. We were, unfortunately, unable to use this information to boost the performance of our system this time, but it is something that should be taken into consideration and explored further when working with offensive, hate, or toxic speech detection.

	pos	*neg*	*ang*	*ant*	*dis*	*fea*	*joy*	*sad*	*sur*	*tru*
NOT	51.476	79.136	30.028	17.623	29.685	47.055	16.588	25.843	10.753	26.785
OFF	68.640	110.350	41.481	23.474	37.543	58.420	22.374	46.046	9.190	26.838

Table 5.21: Normalized emotion word distribution in Danish offensive and non-offensive messages.

We speculated that the reason including this information did not improve the accuracy of the classifier was due to the small size of the dataset. Even though offensive messages were more emotional, non-offensive messages also contained emotion-associated words and due to the large class imbalance, the effect in this case was negligible. With this paper (VIII) we showed that further pre-training and fine-tuning of BERT, specifically with language-specific models, improves accuracy significantly.

5.5.2 The Emotive Rhetoric of Finnish Political Party Manifestos

A second practical study of sentiments and emotions was conducted with political party manifestos as the main dataset. This time my approach was purely lexical. This was due to a few reasons: (1) most sentiment analysis studies in political science, are conducted using lexical approaches[19], and (2) as we had neither the budget or time to annotate the

[19]Nearly all that employ machine learning have been published by computer scientists and not political scientists.

party manifestos manually, a lexical approach was the most feasible one.

For paper IX, where I look at differences in emotion word distribution and emotion word intensities in Finnish political party manifestos, a very similar approach was used as in paper V. However, for the evaluation, rather than cosine similarity as in paper V, we used multilevel linear regression to gain an objective score of both significance and similarity. For the same paper, in addition to SELF we also used FEIL to gauge the intensity of these words. These were presented as histograms as well as subjected to multilevel linear regression. For the input data, the manifestos were lemmatized using the Turku Neural Parser pipeline on the CSC supercomputer *Puhti*.

Many of our findings did not support some previous findings, specifically about government manifestos being significantly more positive than opposition manifestos. Instead, we found that in terms of emotional content, the party manifestos were very alike and have in recent years included more words labeled as positive than previously. The main finding, however, was that although populist party rhetoric is the same as that of other parties in terms of emotion word distribution and content, the intensity of the words they use is significantly higher for negative words. We used multilevel linear regression to test the significance (see figure 5.6).

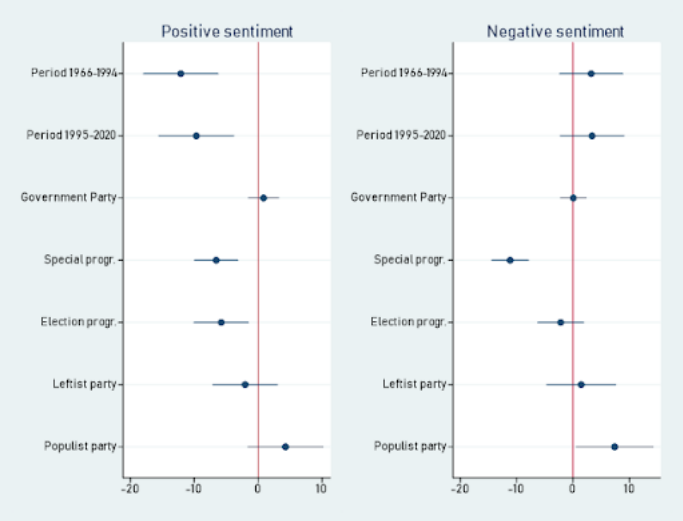

Figure 5.6: Linear regression results

Figure 5.7 shows histograms of emotion words intensities in the party manifestos

of government parties (of which the True Finns have been a part) on the left, and an emotion word intensity histogram for the populist True Finns party manifestos. These histograms too indicate what the linear regression showed: populist parties use more intense words when expressing the same emotions as other parties.

Figure 5.7: Combined coalition government party manifesto intensity histogram (left) and True Finns party manifesto intensity histogram (right)

These findings indicate many interesting details about the emotional content of political rhetoric, especially word choice, however, to me it also shows that lexicon-based methods are still valuable tools for emotion detection in practical applications where the alternative would have been spending a lot of money and time on annotations, neither of which were in supply, or doing subjective close-readings.

5.5.3 Computational Literary Analysis with *Syuzhet*

And finally for literary analysis, we used the SELF lexicon with *syuzhet* (Jockers, 2015b) to automatically create narrative plots of emotional valence for a few classics of Finnish literature. We wanted to test whether the narrative *syuzhet* plots aligned with qualitative analyses of the literature by literary scholars. The findings indicate that the *syuzhet* plots are indeed closely aligned with human interpretations, particularly when used with the SELF lexicon, but that as expected, details and ambiance are missed. Most curiously, the ending of the books tend to be interpreted as positive at worst, and neutral at best even when interpreted by the human, the ending is tragic.

For paper X, where I attempt to combine quantitative and qualitative approaches for

literary analysis in four classic Finnish novels, I use the SELF lexicon to create the same vectors as in some of the previous papers to determine emotion word distribution in the novels. I also incorporate the SELF lexicon to the *syuzhet* R package (Jockers, 2015b). I evaluated the performance of the built-in NRC lexicon in comparison to using the NRC method and the custom method[20], as well as using SELF with the custom method. Furthermore, I analyzed the differences between lemmatized and non-lemmatized texts as the guidelines for *syuzhet* are not clear on this point. As the language I analyzed is Finnish, which is a morphologically rich language — certainly in comparison to English — it stands to reason that lemmatization is more important. This notion is supported by Araque et al. (2019) who recommend lemmatization especially for morphologically rich languages.

All in all, we found that using simple computational methods such as *syuzhet* as a proxy for plot movement, and emotion word distributions can help guide the more subjective qualitative analysis of novels towards a more objective analysis. I would not recommend relying solely on quantitative measures to provide qualitative insights, but quantitative approaches alone can also help us understand literature better if claims are aligned with the chosen methodology. By this I mean that it would be disingenuous to claim solely based on emotion word distributions that a novel is more or less of a specific emotion. There is marked difference in saying *Novel A contains more words associated with anger than novel B when compared to our lexicon* and *Novel A contains more anger than novel B*. The first example is supported by the facts, especially with further specification of limitations. The second one is not. We cannot easily determine whether a lexicon is a better match stylistically or period-wise for one novel or author

[20]The Syuzhet R-package documentation offers very little explanation as to the distinction between these approaches. Indeed, it sounds as if the only difference is that when using the custom method, one is able to use other dictionaries than the built-in ones: the *get_sentiment* function's *lexicon* parameter is described as "A data frame with with[sic] at least two columns named word and value. Works with "nrc" or "custom" method. If using custom method, you must load a custom lexicon as a data frame with aforementioend[sic] columns" (Jockers, 2020, 8). Therefore, it stands to reason that if one extracts the built-in version of the NRC Emotion Lexicon and uses that with the custom method, the results should be identical to using the NRC method. This, however, is not the case, and it remains unclear what the exact distinction between the two methods are. These comparisons are included in the article as a case in point and caveat on using black box tools.

than another.

There are many opportunities for **big data literary analysis**. Using such computational methods, we could compare entire national canons to determine whether one is more depressive than the other, something that has been discussed in e.g. Rossi (2020). But this can only be done with carefully curated and cross-lingually evaluated lexicons.

All in all the study of practical applications indicate that there is room for sentiment analysis and emotion detection in applied sentiment analysis for many different fields outside of NLP. In downstream applications, lexicon-based methods are very useful and can aid in tasks where traditionally close-readings would have been attempted.

Chapter 6

Discussion, Conclusions, and Future Work

The goal of my doctoral studies has been to develop resources and approaches that are widely applicable and models that are explainable and interpretable especially in the context of digital humanities research.

I set out to show that sentiment analysis and emotion detection are useful tools for many real world practical applications and exploratory research in digital humanities. I started out with lexicon-based approaches evaluating emotion preservation between languages and then improved on these lexicons by creating the SELF and FEIL lexicons for Finnish. I also aimed to explore machine learning approaches so I needed annotated emotion datasets for supervised tasks. There was no suitable annotation tool available so I created one and used the feedback from the annotators to design improvements to the tool. Now that I had these annotations, I wanted to examine how well emotions are preserved in translation in order to determine whether annotation projection could be a viable method to creating datasets for different languages. This lead to the creation of the multilingual dataset XED. With these resources, I was ready to test my methods on downstream applications.

6.1 Answers to Research Questions

I attempt to answer the research questions from 1.2 in the following pages. To summarize, what I discovered was that there are research projects where lexicon-based or hybrid methods are suitable due to their high transparency and low initial costs if using existing lexicons. The accuracies of lexicon-based methods are typically lower than those of state-of-the-art neural networks, but interpretability is high, which is important for most traditional fields, including digital humanities, where a qualitative analysis typically follows the quantitative and the research questions are born from a “humanities first” standpoint. Highly optimized engineering solutions often lack transparency and are difficult to analyze and therefore also difficult to be used to understand the underlying problems within the data. For that reason, I focused on interpretable resources, and annotations that enable various kinds of studies on emotional expressions and methods that can detect them.

Typically in computational linguistics or computer science papers on sentiment anal-

ysis and emotion detection the analysis part focuses on an evaluation of the system. In computational humanities and social sciences, the analysis of the results focus on the emotions and how they are triggered and expressed in language. It is therefore much more important for the latter group to have an insight into how the results were obtained and what features are used to detect emotional content.

Neural language models are great at understanding the data in context, i.e. taking sequence and context into consideration, and therefore produce very reliable results. However, supervised approaches require annotated datasets and creating these manually is time-consuming, expensive, and the datasets are typically domain-dependent. Automatic acquisition of annotations works for some domains, like Twitter using hashtags, but such explicit markers of emotions are not available in most domains. As both manually and automatically acquired labels are representative of the source domain first and foremost, accuracies tend to drop when applied to data from other domains. Lexicon-based methods are less domain-dependent (Staiano and Guerini, 2014; Araque et al., 2019) and therefore using one lexicon for many different domains is less problematic than doing the same with machine learning approaches — although lexicons should be fine-tuned for each new domain for better compatibility when possible to increase reliability.

Both lexicon-based and machine learning methods come with some caveats. Chief among these is to take care not to make claims that are not supported by the methodology.

There are a few caveats and considerations that are widely applicable with quantitative methods:

1. Know your data! It is extremely important to understand how the lexicon or dataset was compiled. It is of utmost importance to understand annotator guidelines to be able to draw any conclusions about the results of studies. Some resources are compiled as emotions **evoked** by a word, others as **associated**, yet others as emotions **explicitly expressed** by a word. Sometimes annotators are asked to judge a word unit independently, sometimes in context, and sometimes from the perspective of the writer/speaker, sometimes from the perspective of the reader/listener.

2. Say only what your data actually shows and your methodology supports. When

using simple lexical mapping, you can only claim that you found a certain percentage of emotion-annotated words. To claim anything beyond that, be it author mindset or intentions, or in some cases even that the data **is** more positive or negative simply based on emotion-word distributions, is disingenuous to the data and the method unless further claims are supported by proper empirical evidence.

3. Acknowledge limitations. This can mean the exclusion of valence shifters or genre-hopping, or for some languages using sub-optimal mBERT instead of language-specific BERT. Ideally any method or data that is used is optimized for the particular genre or domain at hand. Using a Twitter dataset to train a machine learning model and testing on student theses, for example, is unlikely to be methodologically sound. It might yield interesting results, but care needs to be taken when making claims about the underlying causes.

Some of these caveats apply to all methods, but whereas the risk with machine learning approaches is not knowing how the results were achieved, i.e. how the model and algorithm made the precise predictions it did, with lexicon-based methods the risk is attributing too much power to individual words and ignoring the influence of context.

Although there are researchers (e.g. Pennebaker et al. 2003) who insist on making assumptions based on simple word occurrence about the author's mindset or personality traits, most consider this idea not substantiated by facts (see e.g. Puschmann and Powell 2018 for a more thorough analysis and criticism). If you know your data and adjust your methods and claims accordingly, the results can speak for themselves and your analysis will be based on facts alone without any unnecessary unsubstantiated claims.

Emotions are generally well preserved in translation, although, what emotions are preserved the best and how certain triggers affect human evaluation differs between languages. It is not always obvious if the problem is with the translation itself, the cultural circumstances of language expression, or the method of evaluation. However, my research has shown that emotions are preserved in translation and therefore both annotation projection and machine translation are viable options.

I showed that datasets and lexicons can be produced for under-resourced languages to gain quality datasets using annotation projection and machine translation. There is a surprisingly small drop in quality, and the results indicate reliable datasets can be created

both with using annotation projection and by employing machine translation and open the door for many new research ventures, including practical solutions for low-resource languages. If both machine translation and parallel corpora exist, this further increases the options available.

The evaluation of the dataset showed that for our data, *surprise* was the most difficult emotion to classify correctly in terms of precision for both humans and computers. However, *disgust* was classified more as *anger* than it was *disgust* and had the worst recall. I would speculate that there are different reasons for the amount of misclassifications for these two emotions. For *disgust*, the reason is likely related to both the similarity and co-occurrence between the emotions of *disgust*, *fear*, and *anger*. For *surprise*, the problem is likely due to *surprise* being both hard to detect in text as it typically relates to a surprising turn of events or new information which is not expressed in text, and it being an emotion that can be both positive and negative. My recommendation is that *disgust* be merged with either *anger* or *fear* or that all three categories be merged, and that surprise be excluded as a class altogether particularly if the modality is text. In chapter 5 and specifically section 5.4.1, I tested this hypothesis outright and showed that combining *anger*, *disgust*, and *fear*, and removing *surprise* increased accuracies significantly.

Emotions are difficult for even the best of models and approaches to determine in text, and it is unclear if this is something that could ever be done as even human annotators disagree about what emotions words, sentences, paragraphs, or even books contain, express, and evoke. In my studies I found that annotator motivation needs to be considered carefully in any annotation task. With intrinsic rewards come intrinsic motivation. I used what I learned from the annotation study to design an improved version of the *Sentimentator* annotation platform, which I then used to create the XED dataset. The XED dataset is probably the major technical contribution of my doctoral research. I showed that our dataset is on par with existing multiclass, multilabel datasets and that datasets in English, which is the language of most datasets, can be translated with free tools such as Google Translate or that using parallel corpora as the source data annotations can be projected to create datasets for under-resourced languages.

Emotion detection has many useful qualities that make it even more suitable to practical applications than traditional sentiment analysis. The ability to discriminate between

different negative and positive emotions is very helpful when determining the difference in rhetoric or narrative devices in various texts. I have shown that emotion detection can successfully be applied in various different downstream applications to objectively show the distribution of emotion words and their relation to the data.

6.2 Scholarly and Practical Contributions

The practical contributions from my doctoral research are the multilingual sentiment and emotion dataset XED, the SELF and FEIL lexicons for Finnish, as well as the *Sentimentator* annotation platform. The XED dataset in particular, but also the SELF and FEIL lexicons to some extent, contribute to increasing the options available for low-resource languages. These have all been thoroughly evaluated and made publicly available on GitHub.

Partly using these resources, I also showed the viability of both lexicon-based methods and machine learning for practical applications. I used machine learning to study offensive speech detection and how emotions are distributed differently in offensive messages as compared to non-offensive. I used lexicon-based methods to study the emotion-laden rhetoric of Finnish political party manifestos and showed that although there were no statistically significant deviations in emotion word distributions between different parties, populist parties used words of higher intensity than other parties. I also established some guidelines for using the *syuzhet* R package in combination with qualitative analyses of affect in Finnish literature and showed that computational methods are a viable support tool for traditional literary analysis.

From a scholarly, or theoretical, point of view I have shown the viability of both annotation projection as well as machine translation of such datasets for emotion detection. This includes studies on emotion preservation in translation which helped prove that annotation projection and specifically emotion detection of projected annotations was a theoretically sound endeavor. Additionally, I also contributed with insights about annotator motivation.

6.3 Future Work

I am looking forward to examining and evaluating the XED dataset more thoroughly. First and foremost, I intend to use the English manual annotations as a base for machine translated datasets in the 109 languages offered by Google Translate in order to create XED-MT. These datasets and the accompanying annotations projected from the English dataset can then be evaluated and the evaluations compared to the projected datasets based on human translations from OPUS. Not only will this show whether MT-based data is better than parallel data for natural language "understanding", it will further help in the creation of new datasets for under-resourced languages. I believe a thorough analysis of the multilingual datasets could potentially help with identifying issues with emotion preservation in translation and guide towards solutions and improvements in information loss between languages.

Furthermore, I have recently submitted a paper on rethinking emotion classification, where I explore possible solutions to some of the issues discussed in this dissertation. Namely, the fact that the emotion theories currently in use focus on psychological theories of emotion that are based on different modalities of human interaction and are not optimized for text. This is a topic that I hope to explore in much greater detail as this is something at the very core of sentiment and emotion analysis for NLP. This future work could also lead to improved guidelines for annotation practices. I have been collaborating with researchers from the University of Turku who are creating their own emotion annotations. There is some interesting overlap between XED and their annotation project that might yield some insights into what exactly makes emotion annotation so difficult. The Turku project is using expert annotators and a different annotation scheme, but have re-annotated parts of XED to match their scheme. This has already revealed some interesting facts, for example, like how *anticipation* is, contrary to previous findings also presented in this dissertation, by far the hardest emotion to categorize into *positive*, *negative* and *neutral*.

I would also like to collect more annotations similar to the original XED data, but much more thoroughly supervise the annotation server-side in order to extract valuable metadata. This would require modifications to the annotation platform, but would yield

insights into annotator behavior and motivation, as well as provide an additional augmentation to the original XED dataset.

As for downstream tasks, I am currently collaborating on several computational literary studies projects. These projects will in the near future likely shed more light on Finnish literature, which has not yet been thoroughly explored using computational methods, but also help create better guidelines for truly interdisciplinary projects.

In my post-doctoral studies I am currently a member of a few different research projects. The first, *Intimacy in data-driven culture* (IDA) at Tampere University, allows me to explore how emotions relate to intimacy and data practices. The second one, *Unconventional communicators in the COVID crisis* (UnCoCo), will allow me to explore the emotions and sentiments of social media posts and compare how people communicated about the crisis during the first and the second wave. This project is linked to the larger *Crisis Narratives* project at Aalto University in which the Finnish Institute for Health and Welfare is also contributing member helping with access to exciting data.

Teaching computational skills, particularly in the digital humanities framework, is something I am very passionate about. I received a small grant from the European Association for Digital Humanities to develop a comprehensive digital humanities focused online Python course. The intent is to develop an interactive course platform where both students and scholars of digital humanities are able to learn progressively more advanced programming topics relevant to their needs and interests, as well as tie together similar existing projects. Hopefully this project will be useful, not only for my own future teaching, but for other teachers and scholars too.

The field of sentiment analysis is constantly moving forward with innovative integrations of hybrid and machine learning methods as demonstrated by recent presentations at influential conferences such as *The 28th International Conference on Computational Linguistics, The 2020 Conference on Empirical Methods in Natural Language Processing*, and *The 58th Annual Meeting of the Association for Computational Linguistics*. New researchers are constantly pushing the field forward by using innovative methods, creating unique datasets, and employing theories from various other fields such as psychology (an excellent example of this is the GoEmotions dataset by Demszky et al. 2020).

In the next few years it is likely that sentiment analysis will become more and more reliable, which in turn will lead to better implementations of systems and methods for downstream tasks in adjacent fields. Perhaps some of these changes will also help pscychologists better define emotions as well.

References

M. Abdul-Mageed and L. Ungar. 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)*, volume 1, pages 718–728, 2017.

M. S. Akhtar, A. Ekbal, and E. Cambria. How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Computational Intelligence Magazine*, 15(1):64–75, 2020.

M. Al-Ayyoub, S. B. Essa, and I. Alsmadi. Lexicon-based sentiment analysis of Arabic tweets. *International Journal of Social Network Mining*, 2(2):101–114, 2015.

F. Ali, D. Kwak, P. Khan, S. R. Islam, K. H. Kim, and K. S. Kwak. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. *Transportation Research Part C: Emerging Technologies*, 77:33–48, 2017.

J. Allwood, N. B. Lindström, M. Börjesson, C. Edebäck, R. Myhre, K. Voionmaa, and E. Öhman. *European intercultural workplace: Sverige*. Gothenburg University, 2007.

C. O. Alm. Characteristics of high agreement affect annotation in text. In *Proceedings of the fourth linguistic annotation workshop*, pages 118–122, 2010.

C. O. Alm. The role of affect in the computational modeling of natural language. *Language and Linguistics Compass*, 6(7):416–430, 2012.

C. O. Alm and R. Sproat. Emotional sequencing and development in fairy tales. In

International Conference on Affective Computing and Intelligent Interaction, pages 668–674. Springer, 2005.

C. O. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics, 2005. URL https://www.aclweb.org/anthology/H05-1073.

A. Andreevskaia and S. Bergler. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 117–120, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1621474.1621496.

O. Araque, L. Gatti, J. Staiano, and M. Guerini. Depechemood++: A bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing*, 2019.

S. Asur and B. A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1, pages 492–499. IEEE, 2010.

Ł. Augustyniak, P. Szymański, T. Kajdanowicz, and W. Tuligłowicz. Comprehensive study on lexicon-based ensemble classification sentiment analysis. *Entropy*, 18(1):4, 2016.

M. Aulamo, U. Sulubacak, S. Virpioja, and J. Tiedemann. OpusTools and parallel corpus diagnostics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789, 2020.

M. Awad and R. Khanna. Support vector machines for classification. In *Efficient Learning Machines*, pages 39–66. Springer, 2015.

G. Badaro, H. Jundi, H. Hajj, and W. El-Hajj. EmoWordNet: Automatic expansion of emotion lexicon using English WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 86–93, 2018.

A. Balahur and M. Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75, 2014.

A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie. Lexicon based feature extraction for emotion text classification. *Pattern recognition letters*, 93:133–142, 2017.

C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 127–135, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

C. Banea, R. Mihalcea, and J. Wiebe. Multilingual sentiment and subjectivity analysis. *Multilingual natural language processing*, 6:1–19, 2011.

J. Barnes, R. Klinger, and S. S. i. Walde. Bilingual sentiment embeddings: Joint projection of sentiment across languages. *arXiv preprint arXiv:1805.09016*, 2018.

L. F. Barrett, M. Lewis, and J. M. Haviland-Jones. *Handbook of emotions*. Guilford Publications, 2016.

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, 2019.

P. S. Bayerl and K. I. Paul. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, 2011.

G. Beigi, X. Hu, R. Maciejewski, and H. Liu. An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment analysis and ontology engineering*, pages 313–340. Springer, 2016.

B. Belainine, A. Fonseca, and F. Sadat. Named entity recognition and hashtag decomposition to improve the classification of tweets. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 102–111, 2016.

E. M. Bender and A. Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://www.aclweb.org/anthology/2020.acl-main.463`.

A. Bermingham and A. Smeaton. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, 2011.

A. Bermingham and A. F. Smeaton. A study of inter-annotator agreement for opinion retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 784–785, 2009.

M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific reports*, 5:10312, 2015.

R. Berwick. An idiot's guide to support vector machines. *Advanced Computer Vision. University of Central Florida*, 2003.

F. Biessmann. Automating political bias prediction. *arXiv preprint arXiv:1608.02195*, 2016.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.

K. Boland, A. Wira-Alam, and R. Messerschmidt. *Creating an annotated corpus for sentiment analysis of German product reviews*, volume 2013/05 of *GESIS-Technical Reports*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim, 2013.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

L. A. M. Bostan and R. Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, 2018.

W. Budiharto and M. Meiliana. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big data*, 5(1):51, 2018.

P. Bühlmann and B. Yu. Boosting with the L 2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

E. Cambria and A. Hussain. Sentic computing. *marketing*, 59(2):557–577, 2012.

E. Cambria, T. Mazzocco, A. Hussain, and C. Eckl. Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space. In *International Symposium on Neural Networks*, pages 601–610. Springer, 2011.

E. Cambria, A. Livingstone, and A. Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.

E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

N. Campbell. Perception of affect in speech-towards an automatic processing of paralinguistic information in spoken conversation. In *Eighth International Conference on Spoken Language Processing*, 2004.

M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171, 2017.

Y. Chen and S. Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, 2014.

K. Cheng, J. Li, J. Tang, and H. Liu. Unsupervised sentiment analysis with signed social networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

R. R. Cornelius. Theoretical approaches to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

C. Cortes and V. Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3:369 – 382, 2019.

C. Crabtree, M. Golder, T. Gschwend, and I. H. Indridason. It's not only what you say, it's also how you say it: The strategic use of campaign sentiment, Aug 2018. URL osf.io/preprints/socarxiv/g2sd6.

A. R. Damasio. Descartes' error and the future of human life. *Scientific American*, 271 (4):144–144, 1994.

C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754, 2011.

C. Darwin. *The expression of the emotions in man and animals*. John Murray, 1872.

K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. doi: 10.1145/775152.775226. URL http://doi.acm.org/10.1145/775152.775226.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, 2017.

R. Delgado and X.-A. Tibau Alberdi. Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS ONE*, 14:1–26, 2019. doi: 10.1371/journal.pone.0222916.

D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020. URL https://identifiers.org/arxiv:2005.00547.

L. Deng and J. Wiebe. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1323–1328, 2015.

S. Desai, C. Caragea, and J. J. Li. Detecting perceived emotions in hurricane disasters. *arXiv preprint arXiv:2004.14299*, 2020.

T. Dettmers. Deep learning in a nutshell: Core concepts. *NVIDIA Developer Blog*, Nov 2015. URL `https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts/`.

A. Devitt and K. Ahmad. Sentiment Analysis and the use of extrinsic datasets in evaluation. In *LREC*, 2008.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

M. Elsner. 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, 2012.

G. Eryigit, F. S. Cetin, M. Yanik, T. Temel, and I. Çiçekli. TURKSENT: A sentiment annotation tool for social media. In *LAW@ ACL*, pages 131–134, 2013.

A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of Language Resources and Evaluation (LREC)*, 2006.

C. Felt and E. Riloff. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145, 2020.

J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 363–370, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL `https://doi.org/10.3115/1219840.1219885`.

J. B. Flaes, S. Rudinac, and M. Worring. What multimedia sentiment analysis says about city liveability. In *European Conference on Information Retrieval*, pages 824–829. Springer, 2016.

A. Foolen. The relevance of emotion for language and linguistics. *Moving ourselves, moving others: Motion and emotion in intersubjectivity, consciousness and language*, pages 349–369, 2012.

N. H. Frijda. The laws of emotion. *American psychologist*, 43(5):349, 1988.

K. Ganesan and C. Zhai. Opinion-based entity ranking. *Information Retrieval*, 2011. doi: 10.1007/s10791-011-9174-8.

J. Gao, M. L. Jockers, J. Laudun, and T. Tangherlini. 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. IEEE, 2016.

L. Gillberg. *Fjäril blir till sommarfluga-problematik i översättning från danska till svenska*. Malmö högskola/IMER, 2005.

A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

S. González-Bailón and G. Paltoglou. Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, 659(1):95–107, 2015.

R. González-Ibánez, S. Muresan, and N. Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, 2011.

D. Gorman. Russian Formalism. *A Companion to Literary Theory*, pages 36–47, 2018.

A. Greenhill, K. Holmes, C. Lintott, B. Simmons, K. Masters, J. Cox, and G. Graham. Playing with science: Gamised aspects of gamification found on the online citizen science project-zooniverse. In *GAMEON'2014*. EUROSIS, 2014.

M. Grinberg. *Flask web development: Developing web applications with Python*. " O'Reilly Media, Inc.", 2018.

O. Gunawan and T. Aldridge. Text mining of Scottish post-emergency and training exercise debrief reports. *Health and Safety Executive (HSE) under contract to The National Centre for*, 2018.

M. Haselmayer and M. Jenny. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & quantity*, 51(6):2623–2646, 2017.

R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

Y. He and D. Zhou. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606 – 616, 2011. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2010.11.003. URL `http://www.sciencedirect.com/science/article/pii/S0306457310000932`.

J. Henderson. The unstoppable rise of computational linguistics in deep learning. *arXiv preprint arXiv:2005.06420*, 2020.

A. Hollsten and A. Helle. *Tunnetko kirjallisuutta?: Johdatus suomalaisen kirjallisuuden tutkimukseen tunteiden ja tuntemusten näkökulmasta*. Number 1425 in Suomalaisen Kirjallisuuden Seuran toimituksia. Suomalaisen Kirjallisuuden Seura, 2016. ISBN 978-952-222-742-3.

E. Hovy. Annotation. *Tutorial Abstracts of ACL 2010*, page 4, 2010.

P.-Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, pages 27–35, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1564131.1564137`.

C. Huang, A. Trabelsi, X. Qin, N. Farruque, and O. R. Zaïane. Seq2Emo for multi-label emotion classification based on latent variable chains transformation. *arXiv preprint arXiv:1911.02147*, 2019.

M. Hægermark. *Att översätta från danska – behövs det?*, pages 117–128. Norstedts, 1997.

S. Ilić, E. Marrese-Taylor, J. A. Balazs, and Y. Matsuo. Deep contextualized word representations for detecting sarcasm and irony. *arXiv preprint arXiv:1809.09795*, 2018.

H. Imada. Cross-language comparisons of emotional terms with special reference to the concept of anxiety. *Japanese Psychological Research*, 31(1):10–19, 1989.

S. Isomaa. Kirjallisuus ja moraaliset emootiot. *AVAIN-Kirjallisuudentutkimuksen aikakauslehti*, 3:5–19, 2012.

E. Jablonka, S. Ginsburg, and D. Dor. The co-evolution of language and emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599): 2152–2159, 2012.

M. Jabreel and A. Moreno. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9:1123, 03 2019. doi: 10.3390/app9061123.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

T. Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226, 2006.

M. Jockers. More Syuzhet validation. *Matthew L. Jockers*, 2016.

M. Jockers. Syuzhet 1.0.4 now on CRAN. *Matthew L. Jockers*, 2017.

M. Jockers. Package 'Syuzhet'. *URL: https://cran.r-project.org/web/packages/syuzhet*, 2020.

M. L. Jockers. Some thoughts on Annie's thoughts . . . about Syuzhet. 2015a.

M. L. Jockers. Syuzhet: Extract sentiment and plot arcs from text. 2015b.

A. Jurek, M. D. Mulvenna, and Y. Bi. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(1):1–13, 2015.

P. N. Juslin. What does music express? basic emotions and beyond. *Frontiers in psychology*, 4:596, 2013.

K. S. Kajava, E. S. Öhman, P. Hui, and J. Tiedemann. Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings, 2020.

N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, 2007.

S. Keen. *Empathy and the Novel*. Oxford University Press on Demand, 2007.

D. Keltner, D. Sauter, J. Tracy, and A. Cowen. Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, pages 1–28, 2019a.

D. Keltner, J. L. Tracy, D. Sauter, and A. Cowen. What basic emotion theory really says for the twenty-first century study of emotion. *Journal of nonverbal behavior*, 43(2): 195–201, 2019b.

M. Khodak, N. Saunshi, and K. Vodrahalli. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*, 2017.

C. S. Khoo and S. B. Johnkhan. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511, 2018.

E. Kim and R. Klinger. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*, 2018.

S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, 2004.

S. Kiritchenko and S. Mohammad. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470. Association for Computational Linguistics, 2017.

S. Kiritchenko and S. M. Mohammad. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1095. URL `https://www.aclweb.org/anthology/N16-1095`.

S. Kiritchenko and S. M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

T. Kirkland and W. A. Cunningham. Mapping emotions through time: How affective trajectories inform the language of emotion. *Emotion*, 12(2):268, 2012.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*, 2015.

J. Koljonen, E. Öhman, M. Mattila, and P. Ahonen. Strength and intensity of sentiments and emotions in party manifestos: Finland, 1945 to 2019. *Scandinavian Political Studies*, 1, forthcoming (submitted).

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

M. Landt. Sentiment analysis as a tool for understanding fiction. In *ACM 2010 Annual Meeting*, 2010.

M. Li, E. Ch'ng, A. Chong, and S. See. The new eye of smart city: Novel citizen sentiment analysis in Twitter. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 557–562. IEEE, 2016.

J. Libovický, R. Rosa, and A. Fraser. How language-neutral is multilingual BERT?, 2019.

R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

P. Lison and J. Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the European Language Resources Association*. European Language Resources Association, 2016.

P. Lison, J. Tiedemann, M. Kouylekov, et al. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 2019.

B. Liu. Opinion mining. In *Encyclopedia of Database Systems*, pages 1986–1990. Springer US, 2009.

B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

C. Liu, M. Osama, and A. de Andrade. DENS: A dataset for multi-class emotion analysis. *ArXiv*, abs/1910.11769, 2019.

P. Lohar, H. Afli, and A. Way. Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108:73–84, 2017.

T. Loughran and B. McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

J. J. Louviere and G. G. Woodworth. Best-worst scaling: A model for the largest difference judgments. Technical report, Working Paper, 1991.

P. Lyytikäinen. How to study emotion effects in literature: Written emotions in Edgar Allan Poe's "The Fall of the House of Usher. In *Writing Emotions: Theoretical Concepts and Selected Case Studies in Literature*, pages 247–64. Bielefeld: Transcript Verlag, 2017.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

B. Maia and D. Santos. Language, emotion, and the emotions: The multidisciplinary and linguistic background. *Language and Linguistics Compass*, 12(6):e12280, 2018.

N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133, 2018.

A. Mathur and G. M. Foody. Multiclass and binary SVM classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5 (2):241–245, 2008.

J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

M. K. McNutt, M. Bradford, J. M. Drazen, B. Hanson, B. Howard, K. H. Jamieson, V. Kiermer, E. Marcus, B. K. Pope, R. Schekman, et al. Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, 115(11):2557–2560, 2018.

W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

W. Menninghaus, V. Wagner, J. Hanich, E. Wassiliwizky, T. Jacobsen, and S. Koelsch. The distancing-embracing model of the enjoyment of negative emotions in art reception. *Behavioral and Brain Sciences*, 40, 2017.

Merriam-Webster Online. Merriam-Webster Online Dictionary, 2009. URL `http://www.merriam-webster.com`.

R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

S. Mohammad. #Emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S12-1033`.

S. Mohammad. A practical guide to sentiment annotation: Challenges and solutions. In *WASSA@ NAACL-HLT*, pages 174–179, 2016.

S. Mohammad. Word affect intensities. Presentation at The International Conference on Language Resources and Evaluation (LREC), 2018a. URL https://www.saifmohammad.com/WebDocs/LREC2018-word-emo-talk.pdf.

S. Mohammad and P. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, 2010.

S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. SemEval-2018 task 1: Affect in tweets.

S. M. Mohammad. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018b.

S. M. Mohammad and F. Bravo-Marquez. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada, 2017.

S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

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

S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480 – 499, 2015. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2014.09.003. URL http://www.sciencedirect.com/science/article/pii/S0306457314000880.

S. M. Mohammad, M. Salameh, and S. Kiritchenko. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130, 2016.

L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, 2011.

I. Mozetič, M. Grčar, and J. Smailović. Multilingual Twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

M. Munezero, C. Montero, E. Sutinen, and J. Pajunen. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(02):101–111, apr 2014. ISSN 1949-3045. doi: 10.1109/TAFFC.2014.2317187.

C. Musto, G. Semeraro, and M. Polignano. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. In *DART@ AI* IA*, pages 59–68, 2014.

M. V. Mäntylä, D. Graziotin, and M. Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16 – 32, 2018. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2017.10.002. URL `http://www.sciencedirect.com/science/article/pii/S1574013717300606`.

H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. doccano: Text annotation tool for human, 2018. URL `https://github.com/doccano/doccano`. Software available from https://github.com/doccano/doccano.

H. Narasimhan, W. Pan, P. Kar, P. Protopapas, and H. G. Ramaswamy. Optimizing the multiclass f-measure via biconcave programming. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1101–1106. IEEE, 2016.

T. Nevalainen, T. Vartiainen, T. Säily, J. Kesäniemi, A. Dominowska, and E. Öhman. Language change database: A new online resource. *ICAME journal*, 40(1):77–94, 2016.

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. EmoHeart: Conveying emotions in second life based on affect sensing from text. *Advances in Human-Computer Interaction*, 2010, 2010.

H. T. Ng, C. Y. Lim, and S. K. Foo. A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX99: Standardizing Lexical Resources*, 1999.

S. Ngai. *Ugly feelings*, volume 6. Harvard University Press Cambridge, MA, 2005.

F. A. Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, 2011.

U. Norberg and U. Stachl-Peier. *Translating between closely related languages: Johan Falkberget's Christianus Sextus in Swedish*. Cambridge Scholars Publishing, 2014.

S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM, 2010.

E. Öhman. Teaching Computational Methods to Humanities Students. In *Digital Humanities in the Nordic Countries 2019*. CEUR Workshop Proceedings, 2019.

E. Öhman. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings, 2020.

E. Öhman. SELF & FEIL: Sentiment and Emotion Lexicons for Finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*, forthcoming (submitted).

E. Öhman and K. Kajava. Sentimentator: Gamifying Fine-grained Sentiment Annotation. In *Digital Humanities in the Nordic Countries 2018*. CEUR Workshop Proceedings, 2018.

E. Öhman and R. Rossi. Affect and Emotions in Finnish Literature: Combining Qualitative and Quantitative Approaches. In *Proceedings of the European Association for Digital Humanities Conference (EADH 2021)*, forthcoming (submitted).

E. Öhman, T. Honkela, and J. Tiedemann. The Challenges of Multi-dimensional Sentiment Analysis Across Languages. *PEOPLES 2016*, page 138, 2016.

E. Öhman, M. Pàmies, K. Kajava, and J. Tiedemann. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference of Computational Linguistics (COLING 2020)*, 2020.

E. S. Öhman, J. Tiedemann, T. Honkela, and K. Kajava. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2018.

L. N. Omondi. Dholuo emotional language: An overview. *The language of emotions: Conceptualization, expression and theoretical foundation*, pages 109–887, 1997.

S. K. Pal and S. Mitra. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks*, 3(5):683–697, 1992.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Information Retrieval*, 2 (1-2):1–135, 2008.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. PYKL, B. Gambäck, T. Chakraborty, T. Solorio, and A. Das. SemEval-2020 Task 9: Overview of sentiment analysis of code-mixed tweets, 2020.

A. Pavlenko. Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism*, 11(2):147, 2008.

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

J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`.

R. W. Picard. Affective computing MIT press. *Cambridge, Massachsusetts*, 1997.

R. Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1: 3–31, 1980.

R. Plutchik. A general psychoevolutionary theory. In P. Ekman and K. Scherer, editors, *Approaches to Emotion*, pages 197–220. Lawrence Erlbaum Associates, 1984.

R. G. Pontius Jr and M. Millones. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429, 2011.

S. Poria, E. Cambria, and A. Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.

S. Poria, E. Cambria, D. Hazarika, and P. Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*, 2016.

C. Puschmann and A. Powell. Turning words into consumer preferences: How sentiment analysis is framed in research and the news media. *Social Media+ Society*, 4(3): 2056305118797724, 2018.

M. Pàmies, E. Öhman, K. Kajava, and J. Tiedemann. LT@Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT? In *Proceedings of the 14th International Workshop on Semantic Evaluation*, 2020.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

J. R. Ragini, P. R. Anand, and V. Bhaskar. Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 42:13–24, 2018.

J. Ramteke, S. Shah, D. Godhia, and A. Shaikh. Election result prediction using Twitter sentiment analysis. In *2016 international conference on inventive computation technologies (ICICT)*, volume 1, pages 1–5. IEEE, 2016.

J. Reilly and L. Seibert. Language and emotion. *Handbook of affective sciences*, pages 535–559, 2003.

V. P. Rosas, R. Mihalcea, and L.-P. Morency. Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, 2013.

R. Rossi. *Alkukantaisuus ja tunteet — Primitivismi 1900-luvun alun suomalaisessa kirjallisuudessa*. Number 1456 in Suomalaisen Kirjallisuuden Seuran toimituksia. Suomalaisen Kirjallisuuden Seura, 2020.

T. Ruokolainen, P. Kauppinen, M. Silfverberg, and K. Lindén. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54(1):247–272, Aug 2019. ISSN 1574-0218. doi: 10.1007/s10579-019-09471-7. URL `http://dx.doi.org/10.1007/s10579-019-09471-7`.

M. Salameh, S. Mohammad, and S. Kiritchenko. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 conference of the North*

American chapter of the association for computational linguistics: Human language technologies, pages 767–777, 2015.

A. E. Samy, S. R. El-Beltagy, and E. Hassanien. A context integrated model for multi-label emotion detection. *Procedia Computer Science*, 142:61 – 71, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.10.461. URL http://www.sciencedirect.com/science/article/pii/S1877050918321628. Arabic Computational Linguistics.

D. Santos and B. Maia. Language, emotion, and the emotions: A computational introduction. *Language and Linguistics Compass*, 12(6):e12279, 2018. doi: 10.1111/lnc3.12279. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12279. e12279 LNCO-0697.R3.

F. S. Satthar. *An inheritance-based lexical approach to sentiment analysis*. PhD thesis, University of Brighton, 2018.

A. Scarantino. The philosophy of emotions and its impact on affective science. In L. F. Barrett, M. Lewis, and J. M. Haviland-Jones, editors, *The handbook of emotions*, chapter 1, pages 3–48. Guilford Publications, 2016.

C. Scheible and H. Schütze. Unsupervised sentiment analysis with a simple and fast Bayesian model using Part-of-Speech feature selection. In J. Jancsary, editor, *Proceedings of KONVENS 2012*, pages 269–273. ÖGAI, September 2012. URL http://www.oegai.at/konvens2012/proceedings/41_scheible12w/. PATHOS 2012 workshop.

N. Scheper-Hughes. The Margaret Mead controversy: Culture, biology and anthropological inquiry. *Human Organization*, 43(1):85–93, 1984.

K. R. Scherer and H. G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66 2:310–28, 1994.

A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10, 2017.

B. Scholkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

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

H. A. Schwartz and L. H. Ungar. Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1):78–94, 2015.

D. Sculley and G. M. Wachman. Relaxed online SVMs for spam filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 415–422, 2007.

K. Shimoda, M. Argyle, and P. R. Bitti. The intercultural recognition of emotional expressions by three national racial groups: English, Italian and Japanese. *European Journal of Social Psychology*, 8(2):169–179, 1978.

E. Shouse. Feeling, emotion, affect. *M/c journal*, 8(6):26, 2005.

G. I. Sigurbergsson and L. Derczynski. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA, 2020.

V. Skirbekk, P. Connor, M. Stonawski, and C. P. Hackett. *The future of world religions: Population growth projections, 2010-2050*. Pew Research Center, 2015.

H. Sklar. *The art of sympathy in fiction : Forms of ethical and emotional persuasion*. Linguistic Approaches to Literature. John Benjamins Publishing Company,

2013. ISBN 9789027233509. URL http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=547591&site=ehost-live&scope=site.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

K. Song, S. Feng, W. Gao, D. Wang, L. Chen, and C. Zhang. Build emotion lexicon from microblogs by combining effects of seed words and emoticons in a heterogeneous graph. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 283–292, 2015.

J. Staiano and M. Guerini. Depechemood: A lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.

P. J. Stone, D. C. Dunphy, and M. S. Smith. *The General Inquirer: A computer approach to content analysis.* MIT press, 1966.

M. Straka, J. Hajič, and Straková. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France, May 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, 2007.

C. Strapparava and R. Mihalcea. Annotating and identifying emotions in text. In *Intelligent information access*, pages 21–38. Springer, 2010.

C. Strapparava, A. Valitutti, et al. Wordnet Affect: An affective extension of wordnet. In *LREC*. Citeseer, 2004.

L. Y.-F. Su, M. A. Cacciatore, X. Liang, D. Brossard, D. A. Scheufele, and M. A. Xenos. Analyzing public sentiments online: Combining human-and computer-based content analysis. *Information, Communication & Society*, 20(3):406–427, 2017.

P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9(4):483–496, 2001.

E. Sulis, D. I. H. Farías, P. Rosso, V. Patti, and G. Ruffo. Figurative messages and affect in Twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132–143, 2016.

A. Swafford. Problems with the Syuzhet package. *Anglophile in Academia: Annie Swafford's Blog*, 2015.

J. Swafford. Messy data and faulty tools. *JSTOR*, 2016.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

J. Tiedemann. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, 2014.

R. Tokuhisa, K. Inui, and Y. Matsumoto. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 881–888. Association for Computational Linguistics, 2008.

S. S. Tomkins. *Affect imagery consciousness: Volume I: The positive affects*, volume 1. Springer publishing company, 1962.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*. Citeseer, 2010.

P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

K. van Es, M. Wieringa, and M. T. Schäfer. Tool criticism: From digital methods to digital methodology. In *Proceedings of the 2nd International Conference on Web Studies*, pages 24–27. ACM, 2018.

C. J. Van Lissa, M. Caracciolo, T. van Duuren, and B. van Leuveren. *Difficult Empathy-The Effect of Narrative Perspective on Readers' Engagement with a First-Person Narrator*. DIEGESIS 5.1, 2018.

A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*, 2019.

G. I. Webb, J. R. Boughton, and Z. Wang. Not so naive Bayes: Aggregating one-dependence estimators. *Machine learning*, 58(1):5–24, 2005.

T. Widmann. How emotional are populists really? *Factors Explaining Emotional Appeals in the Communication of Political Parties (December 1, 2019)*, 2019.

J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer, 2005.

A. Wierzbicka. *Emotions across Languages and Cultures: Diversity and Universals*. Cambridge University Press, 1999.

M. Wilkens. Digital humanities and its application in the study of literature and culture. *Comparative Literature*, 67(1):11–20, 2015.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

D. S. Wirz. Persuasion through emotion? An experimental test of the emotion-eliciting nature of populist communication. *International Journal of Communication*, 12:25, 2018.

S. Wu and M. Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *arXiv preprint arXiv:1904.09077*, 2019.

S. Wu and M. Dredze. Are all languages created equal in multilingual BERT? *arXiv preprint arXiv:2005.09093*, 2020.

X. Yang, S. Obadinma, H. Zhao, Q. Zhang, S. Matwin, and X. Zhu. SemEval-2020 Task 5: Counterfactual recognition, 2020.

D. Yarowsky, G. Ngai, and R. Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. Technical report, JOHNS HOPKINS UNIV BALTIMORE MD DEPT OF COMPUTER SCIENCE, 2001.

J. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1137. URL `https://www.aclweb.org/anthology/D18-1137`.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval2019). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*, 2019a.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1415–1420, 2019b.

M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and c. Çöltekin. SemEval-2020 Task 12: Multilingual offensive lamnguage identification in social media (OffensEval 2020). In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval)*, 2020.

H. Zhang. Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02):183–198, 2005.

L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Chapter A

Appendices

A.1 Copyrights of Introduction and Articles

This introduction (chapters 1-6) is published under *Creative Commons Attribution–NonCommercial-NoDerivs 3.0 Unported*[1] licence, which can be found in Appendix A.3.

A.2 Original Articles

This electronic publication does not contain reproductions of the original articles. The viewers of this electronic version of the dissertation should obtain the articles from the University of Helsinki researchportal[2], Google Scholar[3], or other academic databases.

A.3 Licence for chapters 1-6

This text is *Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported*.[4]

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES

[1] `https://creativecommons.org/licenses/by-nc-nd/3.0/`

[2] `https://researchportal.helsinki.fi/en/persons/emily-%C3%B6hman`

[3] `http://scholar.google.com`

[4] `http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode`

NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN ”AS-IS” BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE. License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE (”CCPL” OR ”LICENSE”). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

   (a) **“Adaptation”** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image (”synching”) will be considered an Adaptation for the purpose of this License.

   (b) **“Collection”** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, con-

stitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License.

(c) **"Distribute"** means to make available to the public the original and copies of the Work through sale or other transfer of ownership.

(d) **"Licensor"** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

(e) **"Original Author"** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

(f) **"Work"** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to ge-

ography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

(g) **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

(h) **"Publicly Perform"** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

(i) **"Reproduce"** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

(a) to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections; and,

(b) to Distribute and Publicly Perform the Work including as incorporated in Collections.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats, but otherwise you have no rights to make Adaptations. Subject to 8(f), all rights not expressly granted by Licensor are hereby reserved, including but not limited to the rights set forth in Section 4(d).

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

(a) You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested.

(b) You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

(c) If You Distribute, or Publicly Perform the Work or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution (”Attribution Parties”) in Licensor’s copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work. The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Collection, at a minimum such credit will appear, if a credit for all contributing authors of Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

(d) For the avoidance of doubt:

i. **Non-waivable Compulsory License Schemes.** In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

ii. **Waivable Compulsory License Schemes.** In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License if Your exercise of such rights is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(b) and otherwise waives the right to collect royalties through any statutory or compulsory licensing scheme; and,

iii. **Voluntary License Schemes.** The Licensor reserves the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License that is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(b).

(e) Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK,

EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

   (a) This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

   (b) Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

(a) Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

(b) If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

(c) No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

(d) This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

(e) The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons Notice

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark ”Creative Commons” or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons’ then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of this License.

Creative Commons may be contacted at `http://creativecommons.org/`.