Modeling Emotion across Languages, Label Formats, and Linguistic Levels

DISSERTATION

im Fach Computerlinguistik
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

eingereicht von Sven Büchel, B.A.

Gutachter:

1. Prof. Dr. Udo Hahn, Friedrich-Schiller-Universität Jena
2. Prof. Dr. Lyle Ungar, University of Pennsylvania
3. Prof. Dr. Veronique Hoste, Universiteit Gent

Tag der mündlichen Prüfung: 25. Oktober 2022
Gesamtprädikat: *summa cum laude*

Abstract

Language-based emotion analysis finds itself in a paradoxical situation. In the past decades, a plethora of datasets have been created, covering diverse aspects of natural language and affective states. However, the considerable volume of resulting gold data is scattered across many design decisions in dataset creation, acting as sources of heterogeneity. Some of these sources of heterogeneity are common throughout NLP (e.g., different natural languages and language registers, as well as linguistic units of different sizes such as words, sentences, and texts). Others are specific to emotion analysis, like the myriad of different label formats (the choice for a particular set of emotion target variables and their respective value ranges), and the selection of a particular viewpoint (e.g., reader vs. writer emotion). Due to this heterogeneity, too often it is impossible to compare or merge existing data from different sources. Conversely, if researchers or practitioners require data that meet a specific set of design decisions, it is unlikely that a suitable dataset already exists. This stands in contradiction to the—in principle—large total volume of emotion gold data.

Still, this heterogeneity is empirically adequate and valuable from an application perspective. Thus, the solution to the heterogeneity problem cannot be to simply reduce the number of possible design choices through community-wide consensus. Rather, what is needed is a delicate balance between fostering the diversity of emotion data and developing new methods to tackle the resulting comparability issues. This interplay between diversity and comparability of emotion data is the focus area of this thesis and the seven studies compiled within it. Some of these studies contribute to data diversity by introducing new datasets and methodologies for the annotation and modeling of complex emotion label formats. Others present methods against incomparability, in particular by transferring knowledge between languages and label formats.

The larger vision behind this dissertation is to arrive at a research landscape where diversity and comparability no longer act as antagonists and instead every new sample of annotated data, regardless of the specifics of its annotation design, benefits the endeavor of emotion analysis as a whole. While NLP is still far from achieving this goal, the presented research results, culminating in the establishment of an “emotion interlingua” in the final study, constitute a significant step in this direction.

Acknowledgments

Writing a dissertation is exhausting, especially so when one's field of work is as much in flux as computational linguistics is right now. At times, it feels like navigating a vast ocean in a tiny vessel in stormy weather. With this image in mind, I feel extremely grateful for every piece of guidance and support I have received over the past years.

First and foremost, I want to thank my advisor Prof. Dr. Udo Hahn for contributing to this dissertation in countless ways. He put trust in my sometimes unusual project ideas, fostered my curiosity and scientific rigor, and made my PhD program, along with research visits and internships, possible. Most importantly though, he always gave me the time necessary to pursue the highest standards in my work. Likewise, I am deeply grateful to Prof. Dr. Lyle Ungar for giving me such a warm welcome to his research group as well as sharing his insights into machine learning and the world in general with me. I look back fondly on my time in Philadelphia and all the knowledge and fascinating insights I acquired there.

Next, I would like to thank my colleagues in Jena and across the world for the uncountable inspiring conversations and discussions. In particular, I would like to express my gratitude towards my former student assistants Tinghui Duan and Susanna Rücker for making my PhD experience more engaging and helping me trust in my teaching abilities. I would also like to extend my sincere thanks to my proofreaders Johannes Hellrich, Erik Fäßler, Luise Modersohn, Tabitha Groß, and, again, Susanna Rücker, who helped me tremendously in improving the quality of the manuscript.

Finally, I would like to express my deepest gratitude towards my friends and family. Knowing that it is not always easy, thank you so, so much for bearing with me!

Contents

List of Tables

List of Figures

Part I

Report on the Dissertation Project

1 Introduction

To experience emotions is fundamental to human nature. They shape our view of the world and our behavior in it, ranging from day-to-day encounters on the street to large-scale economic decision making (Sanfey et al., 2003; Nezlek et al., 2008). A phenomenon of such impact and ubiquity, it is no wonder that emotion has caught the interest of various academic disciplines over the years, with the involvement of modern sciences dating back at least to Darwin (1872/1998). Most foundational perhaps are research activities conducted in the discipline of psychology, e.g., on elicitation mechanisms, structure, and behavioral consequences of emotion (Scherer, 2000). Other disciplines focus on the impact of emotion on economic decision making (Sanfey et al., 2003; Andrade and Ariely, 2009), physical and mental health (DeSteno et al., 2013; Hu et al., 2014; Evers et al., 2018; Guo et al., 2019; Clobert et al., 2020), or the interplay between an individual's feelings and societal structures (Hochschild, 1983; Stets and Turner, 2006; Bericat, 2016). Still others examine which linguistic means (e.g., prosody, lexical choice, or syntax) can be used to express or elicit emotion in the languages of the world (Majid, 2012).

In contrast to these research endeavors that focus primarily on explanation and description, computer science approaches emotion more frequently from an *engineering* perspective, aiming to develop software systems that can fulfill certain affect-related tasks. Here, the study of emotion is mainly situated within the field of artificial intelligence (AI) and is tightly connected to the research program of affective computing (Picard, 1997). Still, different sub-fields of AI, such as computer vision, robotics, and of course natural language processing (NLP; or computational linguistics), pursue their own lines of emotion research in relative independence of each other.

In NLP, interest in affective states has been growing rapidly, beginning around the turn of the millennium. Early research in this field had a strong focus on distinguishing between positive and negative words or statements, a linguistic property referred to as *semantic orientation* or *polarity* (Hatzivassiloglou and McKeown, 1997; Pang et al., 2002; Turney, 2002; Turney and Littman, 2003; Pang and Lee, 2005). Those research activities[1] marked the beginning of a new area of work, known today as *sentiment analysis* (Liu, 2015). This

[1] Similarly, the distinction between *objective* and *subjective* language, i.e., differentiating between factual statements and expressions of personal opinions or feelings, received much interest in these early years (Wiebe et al., 2004).

early focus on polarity soon broadened to allow for more nuanced distinctions between affective states, in particular varying sets of emotion classes such as joy, anger, or sadness (Liu et al., 2003; Alm et al., 2005; Mishne, 2005; Wiebe et al., 2005; Mihalcea and Liu, 2006; Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007) and later also affective dimensions like valence and arousal (Calvo and Mac Kim, 2013; Paltoglou et al., 2013; Yu et al., 2015). In the past decade, those competing frameworks for emotion description have grown significantly in number and complexity (Bostan and Klinger, 2018; De Bruyne et al., 2020), leading to the following, paradoxical situation.

On the one hand, more expressive annotation schemes help develop models that can capture more facets of emotional meaning. This leads to an empirically more adequate description of emotion in language and thus benefits real-world applications (Bollen et al., 2011; Desmet and Hoste, 2013; Schulz et al., 2013). On the other hand, this proliferation of label formats has led to a severe loss in comparability between individual contributions. Although the total volume of available gold data has grown considerably over the years, it is spread over a vast number of annotation schemes. Consequently, comparing or even merging data from different rating studies is often impossible. Moreover, a large number of prediction models are presented in the NLP literature every year but each of them covers only a small part of the full spectrum of human emotion. This problem appears even more pressing when considering other sources of data heterogeneity, in particular the vast number of existing natural languages, their registers, as well as distinct levels of linguistic analysis, but also different viewpoints of emotion understanding such as writer vs. reader emotion. To resolve this conflict, NLP needs new methods that acknowledge the complexity of human emotion without resulting in a landscape of completely incomparable systems and insurmountable annotation cost.

The present thesis, focusing on written, non-dialogue language, compiles seven studies tackling this dilemma from a variety of angles. This includes (a) data collection for previously underrepresented emotional meaning facets such as valence, arousal, dominance, or empathy, (b) methods for "translating" between different annotation schemes, (c) exploiting additional information present in fine-grained label formats through multi-task learning, (d) crosslingual generation of emotion word ratings, and finally, (e) a method for learning a generalized latent representation of emotion that may serve as a starting point to further unify the field of emotion analysis.

This thesis is structured in three parts, each subdivided into several chapters. The first part, which includes this introduction, gives a high-level summary of the dissertation project. Part II provides copies of the articles submitted for examination. Part III gives background information on the author, fulfilling requirements of the local doctorate regulations. Following the current chapter, the remainder of Part I starts by introducing fundamental concepts and problems of emotion analysis in Chapter 2. This provides the basis for presenting my research results in Chapter 3, where each section briefly outlines one of the seven submitted articles in terms of its motivation, methods, results, and impact. The discussion in Chapter 4 reflects on unresolved issues and opportunities for future work. Chapter 5 concludes this summary.

2 Background

This chapter provides a high-level introduction to emotion analysis, focusing on the usage of fine-grained label formats, setting the stage for the presentation of research results in the following chapter. First, I will lay out the psychological foundations of emotion analysis, i.e., what emotions *are* (§2.1) and how they can be formally represented (§2.2). Then, I will move on to describe how emotions manifest themselves in language (§2.3; §2.4). The following sections address common methodologies for annotating (§2.5) and modeling (§2.6) emotion. The final sections of this chapter report on the state of the art in dealing with the heterogeneity of emotion data, in particular in the form of multilinguality (§2.7) and the diversity of existing label formats (§2.8).

2.1 Emotion and other Affective States

What are emotions and how do they differ from feelings, moods, and other affective states? Unfortunately, we lack a universally agreed upon definition. In psychology, the question "What makes up an emotion?" is strongly debated, partly because different theoretical currents diverge in how much attention they pay to various aspects or components such as facial expressions, subjective feelings, or action tendencies (Scherer, 2000). This lack of agreed upon definitions has also found its way into computer science because different research groups have adopted background knowledge and terminology from different parts of the psychological literature (Munezero et al., 2014).

This thesis adopts the definition by Scherer (2000, p. 138f.) according to which "emotions are episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression, and subjective feeling, but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism." In other words, emotions are short-term response patterns to important events (in our surroundings or our own bodies and minds) that may affect us on multiple levels including bodily reactions, feelings, thoughts, and behavior. Importantly for NLP, this definition implies that emotion *per se* is not a language-centered phenomenon. Yet, it opens up several ways in which language and emotion may interact, such as emotion elicitation (where language utterances act as an important external event), verbal descriptions of subjective feelings, or emotionally-colored verbal response behavior (e.g., cursing, wailing, or rejoicing).

Scherer (2000) also proposed a typology of affective states, distinguishing emotions from moods, interpersonal stances, attitudes, and personality traits (see Figure 2.1) that is gaining popularity within NLP (Jurafsky and Martin, 2021).

Emotion Relatively brief episodes of synchronized responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance (e.g., anger, sadness, joy, fear, shame, pride, elation, desperation).

Mood Diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (e.g., cheerful, gloomy, irritable, listless, depressed, buoyant).

Interpersonal stances Affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation (e.g., distant, cold, warm, supportive, contemptuous).

Attitiudes Relatively enduring, affectively colored beliefs, preferences, and predispositions toward objects or persons (e.g., liking, loving, hating, valuing, desiring).

Personality traits Emotionally laden, stable, personality dispositions and behavior tendencies, typical for a person (e.g., nervous, anxious, reckless, morose, hostile, envious, jealous).

Figure 2.1: Typology of affective states by Scherer (2000, p. 140f.).

Looking at the NLP literature through the lens of this typology, it becomes obvious that contributions in the field of sentiment analysis are actually quite diverse regarding the type of affective state they are concerned with: While there is a large body of work addressing short-lived affective reactions, i.e., *emotions* in the sense of the above definition (e.g., Yu et al., 2015; Sedoc et al., 2017; Kim and Klinger, 2018; Mohammad et al., 2018; Troiano et al., 2019), a major share of the studies addressing sentiment in product reviews or political statements arguably aims at capturing "relatively enduring preferences and predispositions", i.e., *attitudes* (e.g., Pang et al., 2002; Hu and Liu, 2004; Socher et al., 2013; Kiritchenko et al., 2014; Mohammad et al., 2016). Interestingly, the term *mood* is often used in the context of time-series analyses of aggregated, group-level expressions of feelings in large corpora, especially Twitter (Mihalcea and Liu, 2006; Bollen et al., 2011; Harsley et al., 2016). This matches the definition of *mood* in Scherer's typology as "diffuse" and having a "relatively long duration".

Finally, Scherer's typology may also help NLP researchers outside of sentiment analysis to better conceptualize their modeling objectives in relation to other areas. For example, we may understand work that is concerned with detecting online harassment or cyberbullying as searching for particular *interpersonal stances* (Huang et al., 2018; Cheng et al., 2021; Ge et al., 2021). Similarly, there is also a quickly growing body of research on detecting *personality traits* from language use (Gjurkovic and Snajder, 2018; Yamada et al., 2019; Lynn et al., 2020).

2.2 Representing Emotion

After having clarified what distinguishes emotions from other affective states, this section introduces common schemes how emotions can be formally represented in the context of annotation or modeling studies.

The term *emotion label format*, or simply *label format*, will be used to refer to such schemes. A label format consists of a set of *emotion variables* such as Anger, Joy, or Disgust, along with their specific value ranges.[1] While the set of variables determines which emotional meaning facets a particular label format can capture, their value ranges indicate the type of learning problem posed by the label format. To subsume both classification and regression data under this term, the present thesis assumes that labels are encoded in the following way: In classification problems, each of the variables takes either “0” or “1”. Depending on the number of variables and the number of allowed “1s” per label, the problem setting is further specified as binary classification, multi-class-single-label classification, or multi-class-multi-label classification. Conversely, in regression settings, each variable takes a numerical score from some real-valued interval, e.g., $[1, 9]$. See exemplary entries in Table 2.1. Annotation methodologies will be introduced in more detail in §2.5.

Sample	Pol	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis
sunshine	1	8.1	5.3	5.4	4.2	1.2	1.3	1.3	1.2
terrorism	0	1.6	7.4	2.7	1.2	2.9	3.3	3.9	2.5
nuclear	0	4.3	7.3	4.1	1.4	2.2	1.9	3.2	1.6
ownership	1	5.9	4.4	7.5	2.1	1.4	1.2	1.4	1.3

Table 2.1: Exemplary ratings for English words according to nine emotional variables: Polarity vs. Valence, Arousal, Dominance (VAD) vs. Joy, Anger, Sadness, Fear, Disgust (BE5). Polarity follows a binary encoding (“1” encodes positivity, “0” negativity), VAD uses 1-to-9 scales (“5” encodes the neutral value) and BE5 1-to-5 scales (“1” encodes the neutral value). Each of the three variable groups, together with their respective value ranges, constitute a label format. Table adapted from Buechel et al. (2020a).

[1] Names of emotion variables are capitalized from here on to differentiate them from the corresponding everyday concepts, e.g., “Anger” vs. “anger”.

In their choice of a particular set of emotion variables, NLP researchers often follow one of the many approaches designed in the long and controversial history of psychology of emotion (Scherer, 2000; Hofmann et al., 2020). One of the major dividing lines, particularly in computational studies, is the distinction between *discrete* (or *categorical*) and *dimensional* approaches to emotion representation (Calvo and Mac Kim, 2013; Canales and Martínez-Barco, 2014; Bostan and Klinger, 2018; De Bruyne et al., 2020). The discrete approach mainly revolves around the notion of cross-culturally universal, evolutionarily derived *basic emotions* such as the six categories identified by Ekman (1992): Joy, Anger, Sadness, Fear, Disgust, and Surprise. Other theorists, however, have proposed diverging sets of basic emotions (Izard, 1971).

In contrast, the dimensional approach is centered around the notion of *affective dimensions*, independent components that, through their respective combination, *compose* our emotions (Osgood et al., 1957; Russell and Mehrabian, 1977; Bradley and Lang, 1994; Broekens, 2012; Bakker et al., 2014). The most important dimensions are Valence (negative vs. positive) and Arousal (calm vs. excited). These two are sometimes extended by Dominance (feeling powerless vs. empowered). Since Valence roughly corresponds to the concept of Polarity (Turney and Littman, 2003), this thesis henceforth also subsumes studies from classical sentiment analysis under the term “emotion analysis”.[2]

Table 2.1 exemplifies the dimensional vs. the discrete approach to emotion representation. This relationship is further illustrated in Figure 2.2. Table 2.2, finally, lists sets of emotion variables and their abbreviations that are frequently referred to throughout this thesis.

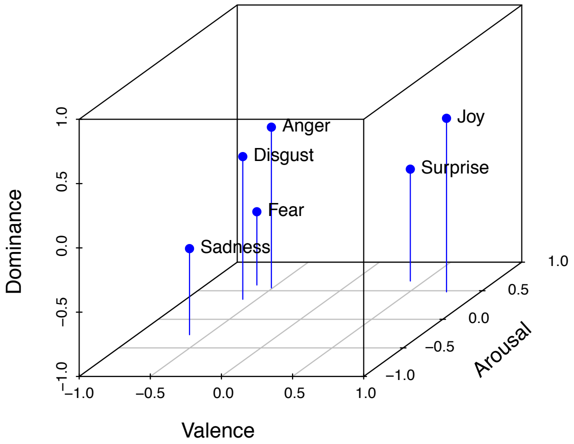


Figure 2.2: Affective space spanned by the affective dimensions of Valence, Arousal, and Dominance, together with the position of six Basic Emotions. Figure adapted from Buechel and Hahn (2016) and originally based on data from Russell and Mehrabian (1977).

[2] In line with the above typology of affective states (§2.1), this only applies to studies that address “emotions” rather than “attitudes”.

Set Name	**Included Emotion Variables**
Polarity	Polarity ($\approx$ Valence)
VA	Valence, Arousal
VAD	VA + Dominance
BE5	Joy, Anger, Sadness, Fear, Disgust
BE6	BE5 + Surprise
Plutchik	BE6 + Anticipation, Awe
Empathy	Empathic Concern, Personal Distress

Table 2.2: Selected sets of emotion variables referenced throughout this dissertation.

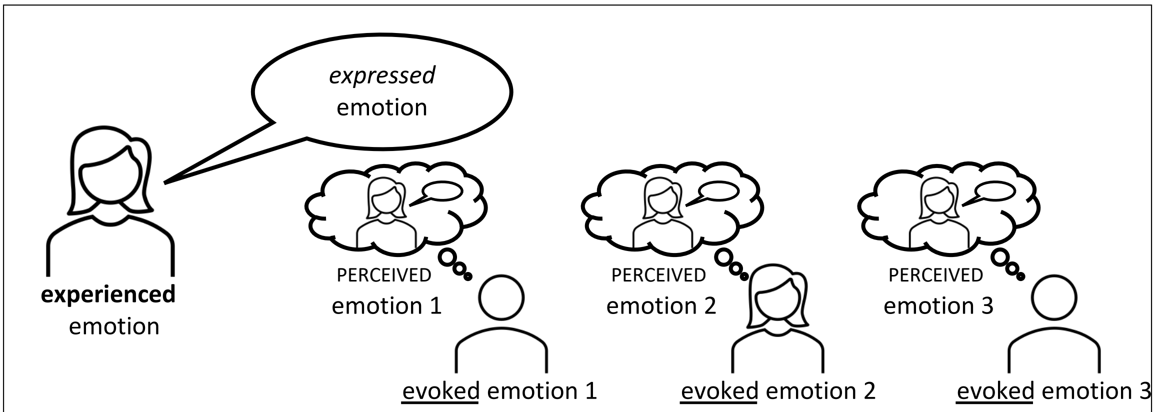
Other theories influential in NLP include Plutchik's (2001) *Wheel of Emotion* (Mohammad and Turney, 2013; Abdul-Mageed and Ungar, 2017; Tafreshi and Diab, 2018; Bostan et al., 2020) and appraisal dimensions (Balahur et al., 2012; Troiano et al., 2019; Hofmann et al., 2020). Yet, frequently studies do not follow any of these established approaches but rather design a customized set of variables in an ad-hoc fashion, often driven by the availability of user-labeled data in social media, or the specifics of an application or domain which requires attention to particular emotional nuances (Bollen et al., 2011; Desmet and Hoste, 2013; Schulz et al., 2013; Qadir and Riloff, 2014; Staiano and Guerini, 2014; Li et al., 2016; Liew et al., 2016; Demszky et al., 2020; Haider et al., 2020). In other cases, researchers focus their work on a specific, often more specialized, emotional meaning facet at a time (Lee et al., 2009; Rouhizadeh et al., 2018).

One of these more specialized emotional nuances, that also plays a larger role in this dissertation, is *empathy*. In particular, I will focus on the two sub-types of empathy proposed by Batson et al. (1987), Empathic Concern and Personal Distress.[3] While Empathic Concern is a warm, compassionate, other-focused feeling for someone in need ("feeling *for* someone"), Personal Distress is a more negative, self-focused affective state in reaction to witnessing someone else's suffering (Buechel et al., 2018). The problem of modeling empathy in language has received much attention, particularly in the speech and spoken dialogue domains (McQuiggan and Lester, 2007; Fung et al., 2016; Alam et al., 2018; Pérez-Rosas et al., 2017). However, research activities on empathy in *written* languages have only recently begun to intensify (Buechel et al., 2018; Rashkin et al., 2019; Sedoc et al., 2020; Zhou and Jurgens, 2020; Guda et al., 2021; Shi et al., 2021; Tafreshi et al., 2021).

[3] See Cuff et al. (2016) for a review of varying definitions and operationalizations of "empathy".

2.3 Viewpoints of Emotion

As outlined in §2.1, language and emotion can interact with each other in multiple ways. For example, language utterances can *evoke* emotion in listeners or readers. Similarly, a speaker can also use language to *express* their own thoughts and feelings. Consequently, units of language may be associated with different emotions depending on the *viewpoint* (or *perspective*) taken during annotation.[4] While NLP researchers have already shown awareness of this in early work (Katz et al., 2007), relatively few studies have been dedicated to this phenomenon specifically. Likewise, there is, to the best of my knowledge, no generally agreed-upon typology of viewpoints. Thus, this thesis proposes to distinguish at least the following ones; see Figure 2.3.



Experienced emotion The emotion actually felt by the writer in the moment of producing an utterance.

Expressed emotion The writer emotion as conveyed by the utterance.

Evoked emotion The reader emotion caused by the utterance.

Perceived emotion The reader's understanding of the writer emotion as induced by the utterance.

Figure 2.3: Proposed typology of the viewpoints of emotion.

Importantly, these four viewpoints do not necessarily agree with each other. For example, writers may choose not to reveal their actual feelings or may fail to properly express themselves, thus causing the Experienced and Expressed emotion to diverge.[5] Also, readers

[4] The remainder of this section uses terminology specific to written language, i.e., "writer" and "reader". The speech bubbles in Figure 2.3 refer to utterances in general, not spoken ones specifically. However, while the work presented throughout this thesis addresses emotion in written language, the concept of viewpoints can of course also be applied to spoken language.

[5] To highlight that they are used in a technical sense, the adjectives "Experienced", "Expressed", "Evoked", and "Perceived" are capitalized when they refer to the respective viewpoint.

may misinterpret the (para-)linguistic clues embedded in an utterance, thus failing to properly understand the writer's feelings. By contrast, if the writer and the reader have a close enough relationship, the latter may still succeed in "guessing" the former's affective state even if not made explicit, e.g., by inferring it from the content of the utterance even if presented as an objective statement. As another example, consider the speech act of threatening someone: Threats are likely to express anger or can even be presented in a friendly voice. Yet, if successful, they evoke fear in the listener. Thus, Expressed and Evoked emotion can diverge as well.

Furthermore, note that the viewpoints differ significantly with respect to what kind of information they capture: An Expressed emotion is a property of a *unit of language*. Gathering ratings for this property thus calls for linguistic annotation methods. In contrast, Experienced and Evoked emotions refer to momentary affective states of *individuals*, thus methods from experimental psychology seem particularly well-suited for collecting such ratings. Lastly, Perceived emotions characterize how individuals see each other, thus neither psychological nor linguistic data collection methods take clear preference (see §2.5 for further details on annotation methodology).

When moving from a one-to-one to a one-to-many communication situation (e.g., traditional mass media or modern social media), additional complications arise from the fact that each utterance now potentially has many readers, again each with their own diverging Evoked and Perceived emotion. Katz et al. (2007, p. 311) give as a simple but strong example the headline "Italy defeats France in World Cup Final". Assuming the author of this statement was a professional journalist from a country other than France or Italy, they probably Experienced neutral emotion. Similarly, since no explicit linguistic cues are given (the utterance is presented as an objective statement), the Expressed emotion is neutral, too. However, the Evoked emotion is likely to vary drastically between readers, depending on their attachment to either the French or the Italian sports team.

Viewpoints of emotion are important for NLP in terms of both annotation and modeling. Yang et al. (2009) show that while the emotion of the writer and the reader of a blog post tend to correlate, both also differ in systematic ways mediated by the topic of the post. In a follow-up study, Tang and Chen (2012) examine which linguistic features are predictive for certain combinations of writer and associated reader sentiment. These two studies illustrate that systematic differences between the viewpoints are important from a modeling perspective since they challenge generalization: If two datasets are annotated with the same label format but different viewpoints, models trained on one dataset are unlikely to generalize well to the other dataset because the relationship between language properties and emotion labels changes according to the viewpoint. Conversely, differences in reader and writer emotion may also be exploited to *increase* modeling performance by addressing both viewpoints with a joint model (Liu et al., 2013; Li et al., 2016).[6] Moreover, the chosen viewpoint of emotion is important for dataset creators as it stipulates how to set up the annotation process, e.g., who to recruit as annotators and how to phrase the respective guidelines. It also influences the resulting annotation quality (Mohammad and Turney, 2013; Buechel and Hahn, 2017a,c; Kajiwara et al., 2021).

While most previous work is relatively easy to subcategorize within the proposed typology (Figure 2.3), some studies use viewpoints that do not fit so neatly into this scheme. Focusing on emotion in poetry, Haider et al. (2020) differentiate between the expressed and the elicited (Evoked) emotion but also introduce the emotion *intended* by the author. Kim and Klinger (2018) and Bostan et al. (2020) are concerned with emotion events *described* by a text, potentially between fictional characters, e.g., "Frodo was angry with Sam" (contrived example). Lastly, Scherer and Wallbott (1994) and Troiano et al. (2019) asked participants to describe a situation from memory in which they felt a particular emotion. Although the resulting data captures the emotion of the writer, it is not necessarily the emotion that the writer felt in the moment of writing. That is, the participants may go through feelings that are similar, yet probably less intense than the original ones when recalling an event.

2.4 Emotion in Language Units of Different Sizes

The last section has described emotion as a property of language emerging on the level of individual utterances in concrete communication situations. However, as this section will lay out, differently-sized units of language have their own way of being "emotional". This thesis will distinguish primarily between the following linguistic levels. Note that I deviate slightly from common linguistic terminology, in an attempt to allow for a more fluent discussion of NLP methodologies.

On the small end of the spectrum, this thesis addresses individual *words*, mostly in the sense of *graphematic words* or *word types*. In contrast, the largest units covered here are long, complete text documents such as novels, business reports or lengthy newspaper articles. I will refer to these as *texts*. Interestingly, what remains of the hierarchy of linguistic units, i.e., phrases, clauses, sentences, as well as short texts, arguably makes up for the majority of NLP research activities. For lack of a better word (and for consistency with the previous section), I will jointly refer to these units of language as *utterances*. See Table 2.3 for an overview of how the terminology of this thesis relates to common linguistic terminology.

[6] In addition to the above studies, there is also a branch of research developing representation formalisms that aim to capture emotional implicatures of certain verbs or event types in relation to different viewpoints (Reschke and Anand, 2011; Deng and Wiebe, 2015; Klenner, 2016; Rashkin et al., 2016). However, this line of work focuses primarily on the polarity of value judgments, thus addressing *attitudes* rather than *emotions* (§2.1).

Having already dealt with utterances above, the remainder of this section describes how emotion emerges from words and texts.

<table>
<tr><th>Terminology of this Thesis</th><th>Usual Linguistic Terminology</th></tr>
<tr><td>words</td><td>words</td></tr>
<tr><td rowspan="4">utterances</td><td>phrases</td></tr>
<tr><td>clauses</td></tr>
<tr><td>sentences</td></tr>
<tr><td rowspan="2">texts</td></tr>
<tr><td>texts</td></tr>
</table>

Table 2.3: Linguistic levels distinguished in this thesis vs. usual linguistic terminology.

The emotion of individual words out of context has received much attention over the years both in psychology and psycholinguistics as well as in NLP (Hatzivassiloglou and McKeown, 1997; Bradley and Lang, 1999; Turney and Littman, 2003; Leveau et al., 2012; Mohammad and Turney, 2013; Warriner et al., 2013; Yu et al., 2015; Li et al., 2017; Buechel and Hahn, 2018c; Mohammad, 2018). Without the framing of a concrete communication situation, the above typology of viewpoints (Figure 2.3) is not fully applicable. For instance, since one examines word types in isolation, there is no particular writer to host the Experienced emotion. Yet, words such as "sunshine", "terrorism", or "hatred" are still clearly emotional, in at least three ways. First, perhaps most obviously, one can conceptualize the emotion of words like "joy" or "hatred" as the emotion they *denote* in terms of their lexical semantics (Majid, 2012). This understanding of word emotion is fundamental for affective extensions to general-purpose lexical resources such as WordNet-Affect (Strapparava and Valitutti, 2004). Second, a word can be understood in terms of the emotional reaction it produces when presented out of context. This corresponds to the Evoked emotion from §2.3. Such an understanding of word emotion forms the basis of many datasets developed by psychologists, which are referred to as affective norm ratings (Bradley and Lang, 1999; Redondo et al., 2007; Kanske and Kotz, 2010; Warriner et al., 2013; Montefinese et al., 2014; Imbir, 2016; Stadthagen-González et al., 2017; see also §2.5). A third sense in which words can have emotional meaning is by (statistical) association. That is, words that appear predominantly in contexts of a certain emotional tone tend to also convey this tone as part of their connotative meaning. Hence, speakers of a language might generally agree that a certain word is *associated* with a particular emotion, without actually *evoking* this feeling when presented in isolation (Mohammad and Turney, 2013). (For example, the word "chocolate" may be generally associated with joy. Yet, reading this word alone may not suffice to lift the spirits.) This understanding of word emotion forms the basis of many computational studies building or extending lexical emotion resources from language usage data (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Staiano and Guerini, 2014; Sedoc et al., 2020).

Compared to the emotions of words, emotions of *texts* are quite challenging in several ways, including conceptually, annotation-wise, and in terms of modeling techniques. Starting with the conceptual issues, the short-lived nature of emotion (§2.1) is at odds with the significant reading time that, say, a novel requires. However, humans still seem able to describe entire books as "happy" or "sad", although what such an attribution truly captures is unclear (this issue is further discussed in §4.1 and §4.4). Putting together this kind of *global* emotion assessment for long texts is also challenging from an annotation perspective. In addition to said conceptual problems, the required reading time makes each individual rating very expensive. Thus, to the best of my knowledge, there is not a single corpus that offers global emotion gold labels for long documents. Close contenders include the dataset by Alm et al. (2005) who annotated complete fairy tales, but sentence by sentence, not globally, and the literary corpus by Liu et al. (2019) which explicitly aims at pushing emotion analysis past the sentence level. Yet, with an average length of only 86 words its samples are still relatively short.

In contrast to this lack of gold data, *applications* of emotion analysis to long documents are actually quite numerous, in particular in the domain of literary texts as well as business and economic reports. This includes many studies that derive global emotion scores despite the above problems (Mohammad, 2011; Mohammad and Yang, 2011; Hajek et al., 2014; Buechel et al., 2016a,b; Goel and Uzuner, 2016; Buechel et al., 2019). An equally important branch of research focuses on the *emotion flow* or *emotion dynamics*, i.e., not computing a global emotion score for the entire document but rather assessing smaller segments with the intention of analyzing how emotions develop over the course of the text (Mohammad, 2011; Kim et al., 2017; Maharjan et al., 2018; Somasundaran et al., 2020). Other studies take an even closer look at long-form texts, e.g., by distinguishing between different channels used to communicate emotion (Are emotions of fictional characters described verbally or conveyed indirectly through the description of facial expressions and bodily reactions?) or looking at relationships between individual character pairs (Nalisnick and Baird, 2013; Kim and Klinger, 2019a,b). Oftentimes, these studies use the derived emotion scores as features in downstream modeling problems such as genre classification or, in the case of literary texts, book success classification (Kim et al., 2017; Maharjan et al., 2018) or financial performance prediction and fraud detection in the case of business reports (Hajek et al., 2014; Goel and Uzuner, 2016).

2.5 Annotating Emotion

After having clarified what emotions are, how they can be computationally represented, and how they manifest in language, the next two sections introduce basic methodologies for annotation and modeling.

Arguably, the most important factor for determining the required annotation methodology is the desired label format (§2.2) stipulating both the variables (or classes) for which human judgments must be collected, as well as the granularity of the respective ratings (i.e., class-based decisions vs. numeric scores). Together, the set of emotion variables and the granularity of their ratings determine the kind of learning problem the resulting data can be used for, i.e., binary classification, multi-class-single-label classification, multi-class-multi-label classification, or (multivariate) regression (§2.2). The set of variables and their granularity also influences the resulting annotation cost: the more variables and the finer the granularity, the more expensive data collection becomes.[7]

Class-based annotations are most often collected from human judges by simply choosing one or multiple emotion categories from a given list. Conversely, numerical annotations are typically collected via rating scales. The number of rating points per scale used in the literature varies. While many psychological studies use scales with five, seven, or nine points (Briesemeister et al., 2011; Riegel et al., 2015; Stadthagen-González et al., 2017), the Stanford Sentiment Treebank (Socher et al., 2013) originally collected their polarity ratings on a 25-point scale before aggregating and binning these annotations into two or five classes. Strapparava and Mihalcea (2007) even used a $[0, 100]$-scale to collect ratings for their AffectiveText dataset. For both categorical and numerical emotion ratings, crowdsourcing has established itself as a viable way to reduce time and cost as well as the reliance on expert annotators (Snow et al., 2008; Mohammad and Turney, 2013; Buechel and Hahn, 2017a; Mohammad, 2018).

Collecting ratings for Valence, Arousal, and Dominance often involves the self-assessment manikin (SAM; Bradley and Lang, 1994). SAM is a set of anthropomorphic cartoon figures that display different levels of emotional intensity per affective dimension (see Figure 2.4). Rating one's feeling with SAM thus comes down to choosing the single most appropriate depiction per row, hence allowing for a language-independent, visual grounding of the meaning of the individual scale points.

Best-Worst Scaling (BWS; Louviere et al., 2015) is a possible alternative to rating scales when gathering numerical emotion annotations, which has recently been introduced to NLP by Kiritchenko and Mohammad (2016). With BWS, raters are given a set of items, typically

[7] In a regression scenario, the number of emotion variables directly corresponds to the number of rating decisions an annotator has to take for each language item. The same is true for a multi-class-multi-label setup. In a multi-class-single-label scenario an annotator may only have to decide for one of the available classes, but this decision becomes more difficult the more contenders there are to choose from.

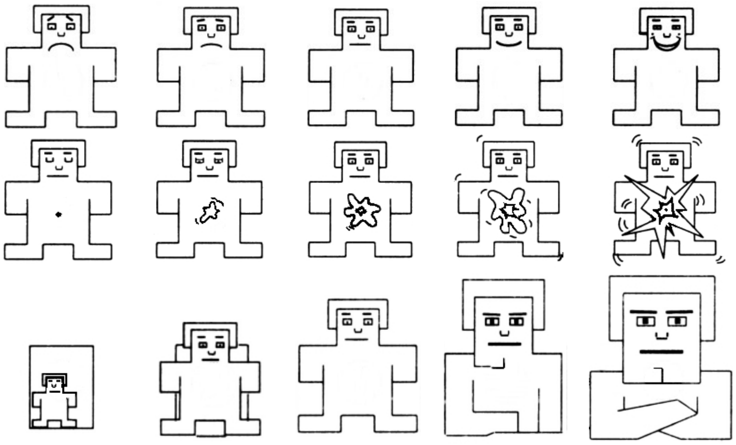


Figure 2.4: The self-assessment manikin (SAM) as presented in Buechel and Hahn (2017a). Rows refer to the affective dimensions of Valence (top), Arousal (middle), and Dominance (bottom). Columns refer to level of intensity from low (left) over neutral (middle) to high (right). Copyright of the original SAM by Peter J. Lang 1994 (Lang, 1980; Bradley and Lang, 1994).

four, and are asked to choose the highest (best) and lowest (worst) item on a particular scale, e.g., Valence. These ordinal judgments are then transformed into metrical ratings using one of multiple available scoring algorithms (Hollis, 2018). While it has been shown that BWS achieves very high *reliability* (see below) for emotion ratings compared to the use of rating scales, i.e., repeated measurements yield similar results (Kiritchenko and Mohammad, 2017), assessing the relative *validity* of BWS ratings is still an area of ongoing research.[8]

A popular alternative to manual annotation, which is particularly well-suited for collecting class-based ratings in the social media domain, is the use of *distant supervision*. In this machine learning setting, training data is generated by heuristically labeling user-generated text, e.g., based on hashtags or emoji (Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017; Felbo et al., 2017).

Class-based annotations are typically evaluated in terms of different variants of the κ statistic (Carletta, 1996), as is the case for many other areas of NLP (Aman and Szpakowicz, 2007; Mohammad and Turney, 2013; Demszky et al., 2020). This family of metrics, however, is not well suited for numerical ratings. Instead, annotation quality of regression datasets is often given in terms of different reliability statistics. *Reliability* is one of the fundamental

[8] Hollis and Westbury (2018) found that BWS ratings for age of acquisition are *more predictive* for behavioral measures, such as lexical decision time, compared to ratings gathered with traditional rating scale methods. The authors take this as an argument for the validity of BWS. However, for Valence, this was *not* the case. Moreover, they found that BWS and traditional ratings display systematic difference, suggesting that both measure slightly different concepts. This finding is also supported by data from Mohammad (2018), who gathered a large VAD lexicon with BWS. While his ratings, again, show very high reliability, the correlation between Valence and Dominance is unusually low. However, even if both rating scales and BWS do measure different things, it is still unclear which of the two measures "the right" thing. Ultimately, the question seems to be whether BWS scores are considered only an economical and reliable way of *approximating* rating scale annotations or whether they count as ground truth in their own right.

quality criteria of empirical research that, at its core, demands that repeated measurements must yield consistent results (Carmines and Zeller, 1979; Hellrich, 2018). Perhaps closest to this idea is the notion of *inter-study reliability* where the repeated measurements stem from completely independent studies. In the context of emotion analysis, this typically means that aggregated gold labels from two distinct datasets are compared, most often in terms of Pearson correlation between instances both datasets have in common (Warriner et al., 2013; Stadthagen-González et al., 2017; Mohammad, 2018; Buechel et al., 2020a). The problem with this approach is that it requires that units of language be annotated with the same label format in different datasets. This is rarely the case, especially for non-English data. Instead, *split-half reliability* can be understood as a way of approximating such agreement through simulation: Using this method, the individual ratings per language unit get randomly assigned into two equally-sized groups. For each of the groups, the individual labels are aggregated *as if they originated from independent studies*. Finally, the agreement between the group aggregates is computed and recorded. This process is repeated, typically 100 times, then averaging the results of the individual runs (Warriner et al., 2013; Buechel and Hahn, 2018a; Mohammad, 2018). Note that, different than inter-study reliability, which can be computed on aggregated ratings alone, split-half reliability requires knowledge of the individual ratings of each annotator. In a modified version of this method, referred to as *leave-one-out reliability*, the groups are not equally-sized. Instead, in each of the runs, one of the raters is in one group and the remaining raters are in the other group (Strapparava and Mihalcea, 2007; Buechel and Hahn, 2017c).

2.6 Modeling Emotion

How does one predict the emotion of a unit of language? Modeling techniques vary drastically depending on the size of the unit they address. This section briefly introduces main lines of model development for the three linguistic levels distinguished in §2.4: words, utterances, and texts.

The prediction of word-level ratings, also referred to as *emotion lexicon induction*, has attracted the interest of NLP researchers since the onset of sentiment analysis (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003). These early studies relied primarily on co-occurrence statistics of word usage extracted from large corpora following an unsupervised approach. Yet, with the wide adaptation of word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Levy et al., 2015; Bojanowski et al., 2017), the dominant methodology used for this problem field shifted. Since then, state-of-the-art contributions typically use pre-trained word representations as input to a supervised machine learning model (Amir et al., 2015; Rothe et al., 2016; Li et al., 2017; Sedoc et al., 2017). In our own work, we could show that this approach can be pushed to a point where it rivals human annotation reliability, when combining high quality pre-trained embeddings with Feed-Forward Networks (FFN; Buechel and Hahn, 2018c).[9]

Modeling the emotion of individual utterances (sentences, tweets, short paragraphs of text) traditionally largely relied on lexical resources, i.e., emotion lexicons. Predictions were mostly based on counting the words in a sample that belonged to a certain emotion class or averaging their scores in case of real-valued labels. Such count-based procedures were often combined with hand-written linguistic rules for handling negation, intensifiers, ordering effects, and similar phenomena (Turney, 2002; Neviarouskaya et al., 2011; Taboada et al., 2011; Hutto and Gilbert, 2014). The next generation of systems still largely relied on lexicons, but these were complemented by other more general-purpose features like n-gram frequencies. Both general-purpose and emotion-specific features were combined and fed into a supervised model to generate a final prediction (Alm et al., 2005; Aman and Szpakowicz, 2007; Mohammad et al., 2013). Thus, such approaches use emotion lexicons mainly as a resource for feature extraction. The arrival of deep learning, thereafter, had a major impact on how emotion in utterances was modeled. In the following years, the dominant approach used word embeddings as input to various deep learning architectures, especially of the convolutional neural network (CNN; Kalchbrenner et al., 2014) and the recurrent neural network (RNN) family, including long short-term memory networks (LSTM; Hochreiter and Schmidhuber, 1997) and gated recurrent unit (GRU; Cho et al., 2014) networks (Wang et al., 2016; Abdul-Mageed and Ungar, 2017; Barnes et al., 2017; Tafreshi and Diab, 2018). Complementary input features, especially derived from lexicons, still yielded some performance gains (Mohammad and Bravo-Marquez, 2017) but overall their importance declined. The most recent development replaces static word embeddings with *contextualized* ones from pre-trained languages models (Akbik et al., 2018; Peters et al., 2018) and, most importantly, transfer learning with transformers (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019; Zhong et al., 2019; Delbrouck et al., 2020).[10]

Regarding text-level emotion prediction, methodological developments are hampered

[9] Complementary to these studies, there is also an extensive body of research trying to enhance the emotional load of word embeddings by incorporating knowledge from lexical resources or emotion-annotated corpora (Tang et al., 2014; Faruqui et al., 2015; Yu et al., 2017; Khosla et al., 2018; Xu et al., 2018). The goal is that computationally derived similarity judgments between words shall not only capture affinity in denotative word meaning (e.g., “cat” and “dog” have greater similarity to each other than to “bridge” because both are animals) but also in connotative meaning (e.g., “sunshine” should become more similar to “chocolate” because both are high-valence words).

[10] This thesis focuses on the emotion of complete utterances in the sense of the viewpoints from §2.3. Other utterance-level problems that derive predictions for sub-sequences, e.g., emotion role labeling, emotion-cause extraction, or emotion span detection (Kim and Klinger, 2018; Xia and Ding, 2019; Oberländer and Klinger, 2020), are considered out of scope. Similarly, modeling the emotion of utterances not in isolation but embedded within dialogue calls for additional methodologies (Poria et al., 2019; Rashkin et al., 2019) and is hence not addressed in this thesis (see Ch.1).

by the lack of suitable gold standard data, i.e., a corpus of long text documents with manually annotated *global* emotion scores (§2.4). Constructing such a corpus is conceptually challenging (What *is* the emotion of a long text document?) and expensive due to the large amount of reading time required. Therefore, existing application studies that derive global emotion scores typically rely on some sort of aggregation step, such as counting or averaging, which propagates affective information from lower linguistic levels (words or sentences for which emotion gold ratings and supervised prediction models are available) up to the text level (Mohammad, 2011; Mohammad and Yang, 2011; Hajek et al., 2014; Goel and Uzuner, 2016). For instance, Mohammad (2011) computed the relative frequency of different classes of emotion words for a collection of novels and fairy tales. This "emotion density" can be seen as a global emotion score which allows to distinguish texts based on their affective characteristics. In other cases, a neural network may be used to predict ratings for the individual sentences of a novel which are then averaged into a text-level score. More abstractly, these simple aggregation steps can be seen as rudimentary approaches to text-level compositionality, i.e., how the overall emotion of a text relates to the emotion of the language units it contains. However, to the best of my knowledge, methods more advanced than these basic arithmetic ones (essentially placing equal weight on all sub-units) have not been developed yet (see discussion in §4.1 and §4.4).

2.7 Emotion and Multilinguality

There are thousands of languages spoken around the world today,[11] each with its own distinct ways of expressing and evoking emotion. This linguistic diversity poses a problem for emotion analysis since in its most successful, supervised form, training data needs to be available for every language of interest, not to mention the multitude of their respective registers and domains. This section outlines two branches of work that can mitigate the resulting, otherwise extreme data requirements by allowing to transfer knowledge across languages.

Firstly, *crosslingual representation learning* is an area of work that is not specific to emotion but can be used to tackle multilinguality in many different NLP areas. Perhaps best known are approaches for learning crosslingual *word* embeddings, that provide vector representations for word types from multiple languages in a shared space (Ruder et al., 2019). Word *sense* embeddings using a multilingual sense inventory such as BabelNet (Navigli and Ponzetto, 2012) can be used in a similar way, yet at an even higher semantic granularity (Camacho-Collados and Pilehvar, 2018). Regarding *contextualized word representations*, the recently proposed multilingual BERT model (MBERT; Devlin et al., 2019) has attracted

[11] `https://www.ethnologue.com/about`; last retrieved on November 29, 2022.

a lot of attention.[12] This transformer model has been pre-trained on many languages simultaneously using a shared, multilingual word piece vocabulary. Consequently, the model embeds utterances from different languages in the same representational space (Pires et al., 2019). These three approaches have in common that they allow training an emotion prediction model in *one* language where gold data is available thereby also enabling inference in *other* languages where gold data is unavailable (Lamprinidis et al., 2021). Crosslingual representation learning thus battles the need to collect training data for every language of interest.

Secondly, there are studies concerned with the *translatability* of emotion. At the center of this line of work stands the question whether a word or an utterance in one language and their equivalents in another language will on average receive similar emotion ratings when annotated with the same methodology. For individual words, psychological affective norm databases provide very strong evidence for this hypothesis (Leveau et al., 2012; Warriner et al., 2013).[13] For utterances, the body of evidence in favor of their (partial) translatability is smaller but expanding (Troiano et al., 2019, 2020).[14] Translatability of emotion ratings is helpful because it suggests that acquiring actual gold data for a target language may be partially surrogated by machine-translating gold data from a source language.[15] This idea has already been implemented in studies, e.g., trying to generate lexical emotion resources for less-resourced languages (see §3.6; Chen and Skiena, 2014; Buechel et al., 2020a; Ramachandran and de Melo, 2020). It also forms the basis for many approaches to crosslingual emotion analysis at the utterance-level (Abdalla and Hirst, 2017; Barnes et al., 2018).

2.8 Sparsity and Incomparability of Emotion Data

Looking back at the large variety of proposed emotion label formats (§2.2), the rise in annotation cost that comes from using such information rich schemes (§2.5), and the distinct viewpoints of emotion (§2.3), combined with the challenges of multilinguality (§2.7), different linguistic levels (§2.4), and other sources of data heterogeneity, it becomes apparent that there is a fundamental problem with emotion analysis.

[12] `https://github.com/google-research/bert`; last retrieved on November 29, 2022.

[13] Mohammad and Turney (2013) make indirect use of this finding by offering machine-translated versions of their well-known NRC Emotion Lexicon: `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`; last retrieved on November 29, 2022.

[14] I am not aware of experimental studies on the translatability of emotion in long texts.

[15] Interestingly, work that models word emotion in historical language stages often relies on the same mechanism. As pointed out in Buechel et al. (2020a), those studies typically come up with a set of seed words that are assumed to have *temporally stable* affective meaning (rather than stability against translation) and then use distributional methods to derive emotion ratings in the target language stage (Cook and Stevenson, 2010; Hamilton et al., 2016; Hellrich et al., 2018; Li et al., 2019).

One way of describing this problem is in terms of *comparability*: Because authors choose different label formats and viewpoints for their work, it is hard to compare annotations from different studies. Existing datasets, even when covering the same language domain, often cannot be merged because of differences in annotation methodology.[16] Also, because annotation cost is quite high, this results in *many small* datasets scattered over various label formats, rather than *few large* ones, which would be more beneficial for achieving high-quality results in prediction and evaluation. As a consequence, existing emotion datasets are often limited in their re-usability. The same is true for emotion analysis software tools: the smaller their coverage of the existing label formats, the fewer researcher will be interested in re-using such software for their own work. Similarly, empirical results, either in the form of model performance or (linguistic, psychological, or cultural) insight, are hard to compare between studies when different label formats are used and distinct viewpoints of emotion need to be taken into account.

Another way of describing this problem is in terms of *data sparsity*. As a thought experiment, imagine compiling all manually annotated emotion data in a single, very large table. The rows of this table would refer to the samples which in turn come from various domains (different linguistic levels, natural languages, and registers thereof) thus representing the heterogeneity of emotion data on the *sample side*. The columns of this table would refer to all the emotion variables that are included in any of the existing datasets, representing the heterogeneity on the *label side*. The individual table cells would then contain the respective ratings (see Table 2.4 for a small-scale illustration). Clearly, this table would be very sparsely populated with only very few language items being annotated multiple times in different formats (e.g., the top two rows of Table 2.4). Most of them are annotated according to only one format, not to mention the virtually infinite number of non-annotated samples that are not included in the table. Now imagine instead that this table was densely populated, i.e., *every* language item that appears in *any* emotion dataset would be annotated for *all* emotion variables. Then, researchers could, per default, develop models and tools, report experimental results and empirical insights according to multiple label formats, in effect mitigating the problems outlined above. However, filling up this table using regular annotation techniques is unfeasible.[17] On top of that, note that the above description of data sparsity did not even cover all sources of heterogeneity but rather left out differences in the value ranges of the ratings (see markers in Table 2.4) and the viewpoints of emotion.

[16] The term *language domain*, or *domain* for short, is used in a broad sense. In this thesis, it refers to combinations of natural languages, genres or registers, and linguistic levels. For example, English words represent one domain while Chinese product reviews constitute another one. In essence, the common usage of the term *domain*, meaning genre or register, is extended to also be applicable across languages and linguistic levels.

[17] Annotating l items for m languages in n label formats results in cubic annotation cost.

Sample	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis
rollercoaster	8.0°	8.1°	5.1°	$3.4^{\square}$	$1.4^{\square}$	$1.1^{\square}$	$2.8^{\square}$	$1.1^{\square}$
urine	3.3°	4.2°	5.2°	$1.9^{\square}$	$1.4^{\square}$	$1.2^{\square}$	$1.4^{\square}$	$2.6^{\square}$
szczęśliwy (pl: “happy”)	$2.8^{\bullet}$	4.0°						
College tution continues climbing				$0^{\blacksquare}$	$54^{\blacksquare}$	$40^{\blacksquare}$	$3^{\blacksquare}$	$31^{\blacksquare}$
A gentle, compassionate drama about grief and healing	$1^{\triangle}$							
喇叭這一代還是差勁透了。 (zh: “This product generation still has terrible speakers.”)	2.8°	6.1°						

Value Ranges: $^{\circ}[1,9]$ $^{\bullet}[-3,3]$ $^{\triangle}\{0,1\}$ $^{\square}[1,5]$ $^{\blacksquare}[0,100]$

Table 2.4: Instances from various datasets described via VAD and BE5 emotion variables. Samples differ in domain (word vs. text, language, register) and label format (covered variables and their value ranges). Translations for non-English samples in lighter shade. Adapted from Buechel et al. (2021).

Importantly, the problem lies not in the diversity of the label formats itself (§2.2). Studies find that those different formats capture affective information complementary to one another (Stevenson et al., 2007; Pinheiro et al., 2017). In line with that, other research shows that, when deriving emotion ratings as an intermediate step for some downstream prediction tasks, different emotion variables turn out particularly useful depending on the application, e.g., stock market prediction, suicide prevention, or crisis management (Bollen et al., 2011; Desmet and Hoste, 2013; Schulz et al., 2013). Consequently, the solution to the outlined sparsity and incomparability problem cannot be for the scientific community to “settle” on a single format (i.e., deliberately reducing the number of columns of the table). Rather, the challenge is to make emotion annotations comparable across formats (i.e., virtually filling up the table) without either giving up representational diversity or having to spend vast amounts of money on data collection.

This challenge, or variations of it, has been recognized by many authors, who in their research approach it from different angles (Hoffmann et al., 2012; Calvo and Mac Kim, 2013; Felbo et al., 2017; Bostan and Klinger, 2018; De Bruyne et al., 2020; De Bruyne et al., 2022). Perhaps the most obvious example are studies that present explicit mappings between different label formats, that act as a kind of annotation projection. This general approach can be implemented in either of two ways. Firstly, it can be studied as a modeling problem in its own right where the goal is to find the best possible mapping function (Buechel and Hahn, 2017a, 2018a; Landowska, 2018). Or, secondly, it can be employed in a two-step-process: Initially, emotion ratings are predicted according to some intermediate label format; then, these ratings are post-processed via label mapping to yield the final predictions according to the desired target format (Calvo and Mac Kim, 2013; Buechel and Hahn, 2016; Hofmann et al., 2020; Zhou et al., 2020; Park et al., 2021). This latter approach can be useful when the only available training data does not follow the required output label format but mapping from the format of the available data to the desired format is easily achieved. Similarly, studies that aim to learn latent representations of emotion that generalize across different annotation schemes show how knowledge can be shared across otherwise incompatible datasets (Felbo et al., 2017; Buechel et al., 2021; De Bruyne et al., 2022).

On a higher level, other areas of work within emotion analysis can also be framed in terms of data sparsity reduction: The development of supervised, monolingual prediction models for the emotion of words or utterances can be seen as a way to transfer affective information *within* a given language from labeled to unlabeled instances (§2.6). Conversely, crosslingual approaches offer a way of transferring knowledge *between* languages (§2.7). Lastly, work on the relationship between various emotion viewpoints may come up with ways to generalize affective ratings given for one viewpoint in order to make them applicable to other viewpoints (§2.3).

The studies compiled for this dissertation sit at the junction of these various problem fields, alleviating sparsity and incomparability of emotion data while simultaneously trying to foster their diversity. The following chapter provides brief descriptions of their individual research contributions, detailing how they fit into this overarching theme.

3 Results

This chapter gives short summaries of the submitted articles; see Table 3.1 for an overview. The individual summaries are structured as follows: They start by describing the broader problem which the respective contribution aims to solve, how this problem is embedded in the overall state of NLP emotion research at the time of conducting this work, and how this particular work is linked to other articles submitted with this thesis. The sections then move on to briefly describe the methods and main results of the respective paper. They conclude by highlighting how this research connects to the various methodological areas outlined in the previous chapter and, in particular, to the overarching theme of this dissertation, data sparsity, comparability, and diversity in emotion analysis.

Reference	Summary	Full	Venue	Rank	Type	Reviews	Accept	Cites
Buechel and Hahn (2016)	§3.1	Ch.7	ECAI	A	long	4	27%	63
Buechel and Hahn (2017a)	§3.2	Ch.8	EACL	A	short	3	24%	176
Buechel et al. (2018)	§3.3	Ch.9	EMNLP	A	short	3	23%	48
Buechel and Hahn (2018c)	§3.4	Ch.10	NAACL	A	long	3	31%	30
Buechel and Hahn (2018a)	§3.5	Ch.11	COLING	A	long	3	38%	24
Buechel et al. (2020a)	§3.6	Ch.12	ACL	A*	long	3	25%	16
Buechel et al. (2021)	§3.7	Ch.13	EMNLP	A	long	3	26%	5

Table 3.1: Articles submitted for examination; with their bibliographical reference, the section in which they are summarized and the chapter that provides their full text, their publication venue and its rank according to the CORE 2021 conference ranking[†], their paper type (long vs. short papers), the number of reviews they received, the acceptance rate of their paper type at the respective venue, as well as their number of citations according to Google Scholar[‡]. ([†]http://portal.core.edu.au/conf-ranks/; [‡]https://scholar.google.com/citations?user=Nwru9iwAAAAJ; both last retrieved on November 29, 2022.)

3.1 Emotion Regression in Affective Dimensions

At the time of preparing our first contribution (Buechel and Hahn, 2016), the research landscape in emotion analysis looked markedly different to today's. While class-based approaches to utterance-level emotion analysis had been popularized some time before

(Alm et al., 2005; Aman and Szpakowicz, 2007), numerical approaches to emotion representation and modeling were still very much in their infancy. These early regression-based studies (i.a., Katz et al., 2007; Neviarouskaya et al., 2011; Staiano and Guerini, 2014) were working almost exclusively on a single corpus with basic emotion annotations: the AffectiveText dataset (Strapparava and Mihalcea, 2007). In contrast, work focusing on affective dimensions, e.g., Valence and Arousal, was almost entirely unrepresented. Lexicon-based methodologies still yielded the state-of-the-art results for emotion regression (Staiano and Guerini, 2014).

Our contribution featured a full review of previous work addressing emotion analysis as a regression problem and discussed conceptual differences between basic emotions and affective dimensions when used in a regression setup. Furthermore, we introduced JEmAS, the first open-source software tool for VAD regression[1] and one of the first emotion analysis systems for VAD in general. JEmAS uses a lexicon-based approach that takes advantage of an affective norm database (§2.4), which originates from psychological research (Warriner et al., 2013), had so far not been used in NLP applications, and is an order of magnitude larger than its well-known predecessor ANEW (Bradley and Lang, 1999). To evaluate our system (which produces predictions in affective dimensions) against previous work (which almost exclusively focused on basic emotions), we trained a second model, acting as a post-processor to our predictions, translating them into the label format of the AffectiveText dataset. Training data for the post-processor was generated by combining pairs of affective norm datasets that have a shared set of entries but use different label formats. In particular, the VAD ratings by Bradley and Lang (1999) and the BE5 ratings by Stevenson et al. (2007) can be combined using one of the two rating sets as labels and the other as input features. To the best of our knowledge, this was the first time that mapping between different label formats has been addressed as a supervised learning problem (§2.8). This approach turned out to work very robustly, so that even after this error-prone post-processing step, our system still achieved state-of-the-art performance for Joy, Anger, and Disgust on the AffectiveText dataset.

This study connects to the topic of this dissertation in three ways: Firstly, it advanced the state of the art in fine-grained emotion analysis on the AffectiveText dataset. Secondly, it is one of the earliest contributions that study emotion analysis with affective dimensions. Through both of these factors, this work contributes to a more nuanced assessment of affective states in language. Thirdly, it is also the first study to propose mapping between emotion label formats as a supervised learning problem, thereby tackling the comparability issue outlined in §2.8.

[1] `https://github.com/JULIELab/JEmAS`; last retrieved on November 29, 2022.

3.2 A Multi-Format Multi-Viewpoint Emotion Corpus

The positive results of the previous study motivated us to further examine the potential of affective dimensions in emotion analysis. The major limitation, however, was the lack of VAD gold data. For example, for the above study, we had only 120 instances with VAD annotations at our disposal (Bradley and Lang, 2007). Such a small sample of data is of course very limiting when developing advanced machine learning models.[2] While two corpora with VA(D) annotations have been published contemporarily to our work, both are still relatively small (less than 3000 instances) and are restricted to online language domains (Preoţiuc-Pietro et al., 2016; Yu et al., 2016).

Consequently, we created our own dataset of genre-balanced, VAD-annotated sentences called EmoBank[3] (Buechel and Hahn, 2017a). The majority of its raw data stems from the written registers of the Manually Annotated Sub-Corpus (Ide et al., 2008, 2010) which in turn is part of the American National Corpus (Ide and Suderman, 2004). Thus, EmoBank not only covers blog posts and newspaper material, but also essays, fiction, and travel guides. We also included the BE6 annotated headlines from AffectiveText (Strapparava and Mihalcea, 2007), so that part of our corpus is annotated twice with different label formats. Moreover, before annotating the full corpus, we conducted a pilot study on the influence of different emotion viewpoints (§2.3) on annotation quality (Buechel and Hahn, 2017c). Since the results were not fully conclusive, we decided to double-annotate the *entire* corpus from two different viewpoints, the Perceived (writer) emotion and a variation of the Evoked (reader) emotion.[4] Ratings were collected via crowdsourcing using the 5-point version of SAM (§2.5). After quality control, the final version comprises over 10K sentences making EmoBank one of the largest emotion corpora at its publication time. In a subsequent analysis, we found that the Evoked reader viewpoint resulted in an overall better annotation quality than the Perceived writer viewpoint. Moreover, similar to the previous study, we conducted a modeling experiment on translating between the included VAD and BE6 labels, achieving performance figures higher than the reported inter-rater reliability of the latter ratings (Strapparava and Mihalcea, 2007). While this result is not easy to interpret (see discussion in §4.3), it clearly indicates that the mapping models presented in Buechel and Hahn (2016) can also be trained efficiently not only on word-level but also on utterance-level data.

[2] However, in a later study, not included in this dissertation, we found that architectures such as CNNs or GRUs can be fitted on as little as 100 data points if well pre-trained word embeddings are used and the number of parameters is kept small (Buechel et al., 2020b).

[3] `https://github.com/JULIELab/EmoBank`; last retrieved on November 29, 2022.

[4] In the interest of quality control, annotators were asked to rate not their own reaction to a piece of text but rather how they thought an average person would answer. This general procedure of asking participants how the general public would respond was later positively assessed by Fujisaki et al. (2017).

In conclusion, EmoBank contributes to the topic of this dissertation in multiple ways. As it is still the largest VAD corpus to this date, it is an asset for advancing fine-grained emotion analysis in affective dimensions, as shown by how frequently the study has been cited (see Table 3.1). So far as I am aware, EmoBank is the first manually annotated multi-viewpoint corpus, thus making possible an array of follow-up studies on the relationship between Perceived and Evoked emotion. Furthermore, its multi-format subsection helps to advance methods of mapping between different label formats. Thus, both its multi-viewpoint as well as its multi-format characteristics contribute to the endeavor of making emotion data more comparable.

3.3 Empathy in Reaction to News Stories

Similar to the above work, the next contribution provides a dataset for previously under-represented emotional meaning facets, here surrounding the notion of empathy. Besides basic emotions and affective dimensions, *empathy* captures an important part of the emotional spectrum which is particularly useful for understanding human interaction, including human-*computer* interaction (§2.2). Previous work in language-based empathy detection was centered around speech data and particularly spoken dialogue while the recognition of empathy in *written* language until recently had been addressed very scarcely. Arguably, one of the main reasons for this is the lack of a publicly available gold standard.

In our study (Buechel et al., 2018), we constructed EmpathicReactions[5], a dataset containing written responses to news articles along with two kinds of empathy annotations. In contrast to the majority of previous work, we used a well-established psychological instrument (Batson et al., 1987) to gather ratings for both Empathic Concern (“feeling *for* someone”), and Personal Distress (“suffering *with* someone”; see §2.2). Ratings were collected using a novel annotation methodology that allowed us to collect reliable judgments from the viewpoint of the Experienced emotion, rather than having to rely on third-party assessments (Expressed or Perceived emotion; §2.3): Participants first read a potentially empathy-evoking news article, then reported their level of Empathic Concern and Personal Distress, before being asked to write a short statement about their thoughts and feelings on the article. The final dataset uses not the newspaper articles themselves but these written responses as samples (1860 after quality control) and the questionnaire ratings as labels. In a subsequent modeling study, we showed that modern deep learning architectures can successfully predict self-reported empathy ratings from the response statements. In terms of performance figures, however, modeling empathy in written language turned out to be a challenging problem, indicating the need for further research.

[5] `https://github.com/wwbp/empathic_reactions`; last retrieved on November 29, 2022.

This work contributes to the advancement of more expressive and fine-grained emotion analysis by providing the first publicly available gold standard for empathy prediction, an important facet of affective meaning that has so far been underrepresented in written language processing. As expected, our dataset has since its publication been used in numerous follow-up studies (e.g., Zhou and Jurgens, 2020; Guda et al., 2021; Shi et al., 2021), including our own work on generating word-level empathy ratings (Sedoc et al., 2020) as well as the WASSA 2021 shared task (Tafreshi et al., 2021). The proposed annotation methodology, which can also be applied to other affective meaning facets than empathy, constitutes one of the first reliable ways to gather ratings for the Experienced writer emotion, thus further advancing viewpoint-aware emotion analysis (§2.3).

3.4 Word-Level Emotion Prediction in Affective Dimensions as Multi-Task Learning

Predicting the emotion of individual words is one of the longest-standing challenges in emotion analysis. Besides being a modeling problem in its own right, allowing to gather valuable linguistic insight, automatically generated word ratings can also be beneficial for detecting affect in larger linguistic units (§2.6).[6] While the importance of lexicons for utterance-level emotion prediction has declined in recent years, they still yield helpful information even for advanced neural network architectures (see §4.1 for discussion). Another advantage of working with word-level ratings is that the respective datasets are available in very high quality for a wide range of languages using consistent acquisition methodologies (§2.4). This makes word datasets highly suited for testing new methodologies, possibly also applicable to larger linguistic units, on many languages in parallel.

In Buechel and Hahn (2018c), we propose a new model for predicting VAD ratings of individual words. Our model[7] is a two-hidden-layer FFN that takes the embedding vector of a word as input, which in turn is given by a pre-trained embedding model. Its main novelty is that its hidden layers are shared between the affective dimensions, only the output layer's parameters (prediction heads) being specific to either Valence, Arousal, or Dominance. Training this model thus constitutes a mild form of multi-task learning (Caruana, 1997). In our experiments, we found that this approach requires *more* training steps before convergence compared to regular single-task learning but ultimately improves model performance by preventing overfitting. Re-implementing a number of earlier systems, we found that our proposed model achieved state-of-the-art results throughout all experimental

[6] To give an example from our own work, for our lexicon-based system JEmAS, we found that using an automatically extended version of a seed lexicon increases performance compared to using the seed lexicon as is.

[7] `https://github.com/JULIELab/wordEmotions`; last retrieved on November 29, 2022.

conditions spanning nine typologically diverse languages. Finally, we found that our model predictions are even competitive to human annotation in terms of split-half and inter-study reliability (§2.5), an observation similar to the one made in Buechel and Hahn (2017a). Again, these findings are further discussed in §4.3.

On the surface, this study advances fine-grained emotion analysis by presenting a new state-of-the-art model for the long-standing problem of predicting individual word emotion. Adapted versions of this model also worked very well in a number of follow-up studies (Buechel and Hahn, 2018a; Buechel et al., 2020a; Sedoc et al., 2020; Buechel et al., 2021). Moreover, since word emotion prediction can be used to induce ratings for previously unrated words, this study also helps to increase the coverage of fine-grained label formats (§2.8). Perhaps equally important though, our work offers a new perspective on the relationship between the expressiveness of a label format and the resulting model performance: Up until this point, the present thesis has primarily described both in an antagonistic way, i.e., higher label informativeness increases annotation cost, which typically results in smaller datasets and thus reduced model performance compared to using simpler label formats (§2.5). But not so in this study—here, we demonstrated that while more informative label formats are obviously more expensive to annotate, they may also *boost* performance by enabling a model to learn more robust representations of emotion, e.g., by multi-task learning. I will come back to these findings in §4.2, where I discuss reasons to choose one label format over another when constructing a new dataset.

3.5 Mapping between Emotion Label Formats

The next study is similar to the previous one in that it uses FFNs to predict emotional word ratings. However, the present contribution focuses not on predicting such ratings from word embeddings but rather revisits the problem of mapping between different label formats (§2.8). This approach already played a subordinate role in Buechel and Hahn (2016) and Buechel and Hahn (2017a), showing promising results. The starting point of this study was the observation that when there are emotion lexicons for more than one format in a language, they are often sized very differently. For instance, when there is a VAD and BE5 lexicon, the former is typically much larger than the latter. In these situations, label mapping can be seen as an alternative to the approach of the previous study. That is, instead of *enlarging* the smaller lexicon (typically the one with BE5 ratings) with word emotion prediction, one may also *translate* the ratings of the larger lexicon to the format of the smaller one. In effect, both approaches yield new ratings for the less-resourced format.

In Buechel and Hahn (2018a), we proposed a two-hidden-layer FFN for label mapping, similar to the one of the above study (Buechel and Hahn, 2018c). Instead of word embedding vectors, our network[8] takes the emotion ratings of a word in one label format (e.g., three real-valued scores for VAD) and outputs the predicted rating in a different format (e.g., five real-valued scores for BE5). In our experiments, we found that this, as an approach to automatic emotion lexicon extension, indeed outperforms embedding-based techniques by a large margin. Our model also outperforms previous approaches to label mapping (Stevenson et al., 2007; Buechel and Hahn, 2016, 2017a,b, 2018b). Moreover, we also proposed a new evaluation methodology for comparing our results to human annotation capacities. Our method addresses the fact that measures of rating reliability (here split-half reliability) are often sensitive to the number of raters, making them difficult to interpret (see discussion in §4.3). Using this method, we found that the reliability of our predicted labels is on par with the reliability of a reasonably-sized group of human annotators. This even held true in a crosslingual setting, where the model was fitted on data from one language but then tested on data from another language. This suggests that the mapping between the VAD and the BE5 format is language-independent.[9] Finally, we used label mapping as a tool to study the relative importance of emotion variables, finding that Dominance is the least important variable for predicting BE5 and Disgust is the least important one for predicting VAD (see discussion on what constitutes a "good" label format in §4.2).

In its attempt to "translate" between different label formats, this study directly targets the issue of incomparability of emotion data (§2.8). Conversely, when applied to automatic lexicon creation, our approach decreases data sparsity by providing an already annotated language item with ratings for an additional label format. While in this study we focused on lexical data again, I emphasize that the presented mapping model is in no way restricted to word ratings but can be applied to utterance-level ratings without any modification.[10] The final two studies of this thesis revisit and extend the notion of "translating" between label formats.

3.6 Creating Massively Multilingual Emotion Lexicons

The next study brings together the results of the two previous ones and extends them through a crosslingual perspective: Firstly, in Buechel and Hahn (2018c), we showed that predicting word-level emotion achieves very high performance and may thus be seen as a way of expanding the coverage of a monolingual emotion lexicon in terms of the number

[8] `https://github.com/JULIELab/EmoMap`; last retrieved on November 29, 2022.

[9] We gathered further evidence for this in a previous study, not included in this dissertation (Buechel and Hahn, 2018b).

[10] Besides Buechel and Hahn (2016) as well as Buechel and Hahn (2017a), results from yet another study not included here also support this claim (Buechel and Hahn, 2017b).

of *word entries* in the same format. Secondly, in Buechel and Hahn (2018a) we found that emotion label mapping produces even more reliable ratings and can thus be understood as a way to extend the coverage of an emotion lexicon in terms of the number of *emotion variables*. Thirdly, previous work had shown that translational equivalents of (especially) words typically receive very similar emotion ratings across different languages (§2.7). That is, emotion ratings are, to some degree at least, invariant against translation.

In Buechel et al. (2020a), we combined these observations by presenting a methodology for generating large, representationally diverse emotion lexicons for any target language imposing only relatively mild data requirements. Starting with a source language emotion lexicon, our method uses label mapping to extend its coverage to more emotion variables, as was done in Buechel and Hahn 2018a. Next, we machine-translate the enriched source lexicon into the target language. We then train a word-level prediction model for the target language using these machine-translated word entries together with their *source* language ratings as (silver standard) training data, following the approach from Buechel and Hahn (2018c). Finally, we predict new ratings for the target language using our newly trained model, replacing and extending our silver standard ratings. This prediction step also mitigates errors introduced during translation. Besides this *generation* procedure, an equally large technical challenge lied in designing a robust *evaluation* scheme. This was due to phenomena linked to lexical ambiguity which may cause knowledge to leak between train and test data.[11] Using our method, we generated emotion lexicons for 91 languages, each of them covering at least 100,000 words.[12] Our evaluation indicates that their quality is about equal to monolingual generation and, again, competitive to human reliability.[13]

As detailed in §2.8, one of the main drawbacks of more complex emotion label formats lies in their increased annotation cost which often limits the coverage of the respective resources, especially when considering that they need to be (re-)created for every language of interest. The summarized study mitigates this issue by expanding the volume of annotated data along multiple axes to include new words, emotional variables, and other languages. It thus efficiently battles the sparsity problem of fine-grained emotion label formats. Nevertheless, this study is still restricted to the word level. Future work may investigate whether similar expansion mechanisms can be applied to larger linguistic units as well—a conjecture for which some pieces of supporting evidence already exist (Troiano et al., 2019, 2020).

[11] For instance, lexical ambiguity on the target-side may result in duplicate entries in the silver standard, because multiple source-side entries translate to the same target-side word. For example, both English *spring* and *feather* translate into the German word *Feder*.

[12] `https://zenodo.org/record/3756607`; last retrieved on November 29, 2022.

[13] In this case, the method for comparison against human reliability presented in Buechel and Hahn (2018a) could not be applied due to data limitations, so that we used inter-study reliability instead. However, the fundamental difficulties of comparing model prediction quality against human reliability still apply (see discussion in §4.3).

3.7 Label-Agnostic Emotion Embeddings

Besides only focusing on the word level, the above work is limited by the fact that, while the proposed method *creates new* ratings, it is less suited for *comparing existing* ones. For instance, if two independent annotation studies gathered ratings for a shared group of language items using different label formats, how could one determine whether these studies agree or perhaps systematically differ in their findings? Thus, in terms of the problem description from §2.8, the above study tackles the *sparsity* of emotion data but only to a lesser extent their *incomparability*. One possible solution may be to apply label mapping to one set of ratings, thus translating them to the format of the other rating set. However, this approach may still be insufficient for the purpose of comparison, since different label formats capture diverging parts of the full emotional spectrum (§2.2, §2.8, Buechel and Hahn, 2018a). Furthermore, in the form presented in Buechel and Hahn (2018a), label mapping only operates on *pairs of label formats*, i.e., a mapping model has a specific input and output format. Thus, a quadratic number of mapping models would be necessary to cover all possible translation directions, not to mention the training data needed to fit those models. While the studies compiled in this dissertation mainly address the BE5 and VA(D) format, the number of label formats in use is actually much larger than that (§2.2). Consequently, instead of relying on label mapping as presented above, I argue that a more desirable solution would be to take into account the fact that, at their core, all label formats try to capture (different facets of) the same underlying quality, human emotion, which is not specific to any particular language domain.

In Buechel et al. (2021), we implemented the above considerations by learning a distributional representation of emotion, so-called emotion embeddings, that generalizes over different label formats, language domains, and model architectures.[14] On the technical level, our method first learns a *multi-way mapping model* which extends over the mapping model presented in Buechel and Hahn (2018a) in that it translates between multiple label formats, rather than only two. It does so via a shared intermediate layer which can be thought of as an “interlingua for emotion”. In a second step, the format-specific output layer of the multi-way mapping model (i.e., its prediction heads) is re-purposed to be deployed on top of existing model architectures for word or utterance level emotion prediction (e.g., the FFN from Buechel and Hahn (2018c) or a BERT model). Applying our proposed training scheme, these existing models learn to embed their respective samples in the common emotion space, and in the process also learn to predict ratings according to additional label formats. As our evaluation shows, these benefits come without any negative impact on prediction quality.

In summary, our method embeds the emotion of language items from various domains in a shared representational space, thus offering a solution to the incomparability problem

[14] `https://zenodo.org/record/5651129`; last retrieved on November 29, 2022.

outlined in §2.8. We consider our emotion embedding technique particularly valuable for future research studying emotional language use across linguistic and cultural barriers. Future work may expand the coverage of the prediction heads to additional label formats and explore their suitability for modalities other than written language. Linking together different research branches presented in this thesis—predicting emotions of individual words and utterances, translating between label formats, and battling sparsity and incomparability of emotion data—, this study constitutes the focal point of my dissertation project.

4 Discussion

This chapter aims at revisiting some of the issues which were raised in the previous sections but not fully addressed by any of the presented studies.

4.1 Do Word Ratings Still Matter?

A large portion of the work presented in the preceding chapter addressed the prediction of word-level emotion ratings, in particular Buechel and Hahn (2018c), Buechel and Hahn (2018a), and Buechel et al. (2020a). While emotional word ratings have inherent value from a linguistic perspective and facilitate the designing of psychological experiments (Monnier and Syssau, 2008; Hofmann et al., 2009), it is also apparent that for most industry applications predictive models for the utterance or even text level (§2.4) are in higher demand.[1] Consequently, one might wonder what practical benefits emotion lexicons have from an application perspective.

An obvious advantage of such lexicons is that they can and have been used to improve the performance of utterance-level prediction systems (§2.6). Historically, sets of word ratings have been one of the central components, obviously, in the lexicon-based approach. However, their importance dwindled as the dominant methodologies changed, first to feature-engineering-based machine learning, then to end-to-end deep learning. Yet, evidence suggests that even today's transformer architectures benefit from additional information provided by lexicons (De Bruyne et al., 2022).[2] Additionally, methods that aim to increase the emotional load of word embeddings, thereby improving downstream performance as well, also often rely on such lexical resources (Faruqui et al., 2015; Yu et al., 2017; Khosla et al., 2018).

Another argument in favor of word ratings is that a lexicon-based approach to emotion analysis is much cheaper, both in terms of required computing and annotation effort, compared to a sophisticated but expensive neural approach (§2.5). Thus, lexicons are

[1] One notable exception are writing assistance applications, which often involve some sort of lexical substitution method, e.g., suggesting to replace part of the words in a document with ones that better fit the intended style (Troiano et al., 2021b).

[2] In a broader sense, integrating emotion lexicons with end-to-end neural network models can be seen as part of a larger research endeavor to link deep learning approaches with symbolic knowledge bases (Sukhbaatar et al., 2015; Logan et al., 2019).

well suited for languages where no training data is available. While recently proposed multilingual language models (such as MBERT; §2.7) are a strong contender for this usage scenario, they still require sufficient amounts of raw data for the target language, giving lexicons a solid use case for under-resourced languages (Tafreshi, 2021). This seems especially true in the light of the presented results on the translatability of word emotions (Buechel et al., 2020a).[3]

Lastly, in the case of text-level emotion analysis, word-based affect scores may also capture information complementary to that of utterance-level models. Recall that models cannot be trained to predict the emotion of long text documents (novels, business reports, etc.) in the regular, supervised fashion due to the lack of gold data and, more importantly, the conceptual difficulties in creating them (§2.4). However, from an application perspective, it may still be necessary to derive a global emotion score from such a long document (§2.6). An obvious way to apply an utterance-level model to a long text document would be to divide the document into multiple slices (e.g., sentences or paragraphs), derive a prediction for every slice, and then average these individual predictions. Note that in doing so, one essentially assigns equal weight to every slice, thus ignoring compositional effects on the text level.[4]

Interestingly, since no gold data is available, the question whether utterance-level neural networks outperform a lexicon-based approach on the text level, as they certainly do on the utterance level, cannot reasonably be answered right now. Instead, one may compare word and utterance approaches based on their predictiveness for downstream modeling problems, e.g., using their respective outputs as features for, say, book sales or bankruptcy predictions (§2.4). As far as I am aware no such study has been conducted yet. However, it seems reasonable to assume that, similarly to how different emotion variables can be particularly useful depending on the given downstream problem (§2.2), word-level emotion may be more predictive than utterance-level emotion for some applications but not for others. In other words, emotion lexicons may provide useful, complementary information for downstream modeling problems.

4.2 How to Compare Emotion Label Formats?

One of the fundamental arguments in favor of more complex representation schemes of human emotion is that they render the output of NLP systems more informative for

[3] The final lexicon version produced by our method (Buechel et al., 2020a) relies on a large-scale target language embedding model and is therefore not easy to generate for severely under-resourced languages. However, our data also show that the machine-translated intermediate versions of the lexicons are of satisfactory quality already. Creating this lexicon version only requires word-to-word translation which should almost always be available when using English as the source language.

[4] This is similar to how a purely lexicon-based system ignores compositional effects on the utterance level.

downstream applications (§2.2): Since application scenarios differ in how much they benefit from a particular emotional nuance, collecting data for a diverse set of label formats is important. However, does this mean that all label formats are “created equal”? Or are some formats objectively better than others? In other words, is there a way to quantitatively compare different label formats? I argue that among the most important properties of a “good” label format are what I will refer to as (high) informativeness and (low) annotation cost.

Perhaps the most obvious way to operationalize the notion of *informativeness* is by looking at downstream predictiveness, i.e., using the output of an NLP system as the input to a secondary modeling problem. Yet again, such an approach is unlikely to produce cut-and-dry results since the outcome will depend on the particular downstream application (§2.2). Another approach to the notion of informativeness is to look at results from label mapping experiments. In Buechel and Hahn (2018a), we conducted an ablation study examining which of the emotion variables of the VAD and BE5 formats are most useful for predicting the other set of variables, respectively. We found that Dominance is the least important variable for predicting BE5 ratings, and Disgust is the least important one for predicting VAD ratings. One may thus conclude that the *gain* in informativeness of the VAD format compared to VA is relatively small, so that using the latter format would be the better choice in most situations. However, the problem with this approach is that it only yields an assessment of informativeness *relative* to another format, in this case BE5. Yet, Dominance may turn out much more important when mapping between VAD and, say, the eight categories by Plutchik (§2.2). Moreover, these results do not necessarily agree with findings from the aforementioned downstream predictiveness. For instance, we have found in several application studies (not included in this thesis) that Dominance is actually *more* important than Valence and Arousal for discriminating between certain text genres or tracing policy change through time (Buechel et al., 2016a,b, 2019). Obviously, these results are not yet fully conclusive, calling for more research in the future. A good place to start may be a systematic review on the importance of different emotion variables for various downstream applications as well as turning to more formal means to study the informativeness of different label formats.

Annotation cost here refers to the necessary monetary compensation to collect a single, individual rating for a unit of language (not an aggregated gold label). As described in §2.5, main influence factors of annotation cost include the number of covered emotion variables as well as the granularity of the annotation (class-based vs. numeric ratings and in the latter case the number of rating points). Moreover, it seems plausible that some emotion variables are inherently harder for humans to annotate than others, causing more cognitive load (e.g., Valence vs. Surprise). Other factors that may influence per-rating cost are the domain and the length of the language samples (e.g., long sentences from legal documents are likely to take longer to annotate than short sentences from TV show transcripts). However, these two factors seem mostly independent of the chosen label format.

Obviously, informativeness and annotation cost of a label format result in a trade-off: the more emotion variables and the finer the granularity, the higher the cost tends to be per rating.[5] To capture how well a label format manages to balance these two contradictory requirements, I propose to use a notion of *annotation efficiency* (similar to how Precision and Recall are combined into the F-Score). To illustrate why this notion is important, imagine you have to create the first gold dataset for a new language domain (no other gold data exists) using a fixed budget. Your subsequent goal is to develop an NLP system with the intention to use its output as an input for a downstream modeling application. Again, this leads to the following trade-off: One could either go for a highly informative label format, resulting in high per-item cost and a small dataset, or one could choose a less informative label format leading to a larger dataset (potentially enabling the training of a more robust model) but with less expressive labels. Which choice of label format will ultimately result in the highest downstream modeling performance? For example, consider the problem of modeling the development of an economic indicator such as the gross domestic product (GDP) of Germany based on emotion indicators derived from German newspaper articles with a fixed budget of 2,000 USD to annotate training data (assuming there would be no German emotion dataset). In terms of the final GPD forecasting performance, would it be better to annotate, say, 10,000 instances with Valence only or 5,000 instances with Valence and Arousal? To the best of my knowledge, the research necessary to answer this question is still very much in its early stages. Importantly, however, the studies compiled in this thesis point out interaction effects between informativeness and prediction quality that may be exploited to improve the annotation efficiency of a label format. In particular, Buechel and Hahn (2018c) show that by making use of multi-task learning, a larger number of emotion variables can lead to better, more robust models, hence counterbalancing the negative effects of having less training data due to higher annotation cost.

In summary, the diversity of existing label formats is advantageous for downstream applications, since different emotional nuances may be more or less important depending on the specific use case. However, this does not mean that all label formats are equally well suited for data collection. Rather, formats should be *quantitatively* compared regarding their annotation cost and their informativeness for a specific downstream modeling problem. This turns the assessment of different label formats into an empirical question calling for future work.

[5] Another important factor that has been omitted for simplicity is the reliability of the ratings. That is, some label formats may be more prone to disagreement between annotators than others and may thus require the collection of more individual ratings per sample to be used as training data. For example, there is strong evidence that Valence and Joy cause much less disagreement between raters than Arousal and Disgust (Buechel and Hahn, 2018a).

4.3 How to Compare Human and Machine Performance?

The studies compiled in this dissertation, on multiple occasions, make comparisons between model performance and human performance, in terms of rating reliability[6], often finding that the former achieved a performance figure about as high as (and sometimes even higher than) the latter (Buechel and Hahn, 2017a, 2018c,a; Buechel et al., 2020a). Surely, these are noteworthy observations, underscoring the high quality of the respective model. Yet, coming up with an accurate interpretation for these findings is not without difficulty.

The most obvious problem is that throughout these studies, different methods for computing human reliability have been used. In particular, we used leave-one-out reliability in Buechel and Hahn (2017a), split-half reliability in Buechel and Hahn (2018a), and inter-study reliability in Buechel et al. (2020a). In Buechel and Hahn (2018c), we relied on both split-half and inter-study reliability (see §2.5 for a description of these measures). The reason for this—perhaps confusing—heterogeneity is that these methods pose very different data requirements. While split-half and leave-one-out reliability require knowledge of the *individual* ratings of an annotation study, inter-study reliability relies on comparing the final, *aggregated* labels from *two independent* studies. However, most dataset creators do not share the individual ratings underlying their gold labels (e.g., out of annotator privacy considerations) and most language items are not annotated in multiple datasets, let alone using the same label format. Consequently, computing reliability measures for third-party datasets is rarely possible. Thus, when comparing model against human performance, most of the time one is limited to whatever reliability figure the respective dataset creators provide.

A more fundamental problem for performance comparison is that the task specifications for humans and machines are not strictly identical. To illustrate what I mean by this, consider the Evoked emotion viewpoint (§2.3): Human raters are given the subjective task to indicate *their own* feelings. The system, on the other hand, is evaluated based on how well it predicts the aggregated response of *many* human raters. However, if the human task and the machine task differ in such an obvious way, how is it possible to compare their performance at all? This question touches on the relationship between annotation *agreement*, annotation *reliability*, and model *performance* which, as I will argue, turns out to be slightly different when annotating emotion compared to more traditional linguistic annotation tasks, e.g., for part-of-speech tags.

[6] I will further elaborate on my understanding of "human performance" and how it relates to "rating reliability" in the course of this section. However, the starting point is the question whether, if the model predictions agree more with human ratings than human ratings tend to agree with each other, does this mean that the predictions can replace actual gold data?

In computational linguistics, the notion of inter-annotator agreement is strongly tied to (different variants of) the κ statistic (§2.5). This family of agreement measures is based on how much more than chance a *pair* of annotators agrees with each other.[7] However, emotion ratings, at least for the Evoked and Perceived viewpoint, are inherently subjective. Hence, a lack of agreement between a particular pair of raters cannot necessarily be taken as an indicator of low annotation quality (Davani et al., 2021; Troiano et al., 2021a). Arguably, this is the reason why reliability measures such as inter-study reliability and split-half reliability—which capture how strongly the aggregated responses of two groups of raters agree with each other—are popular as an indicator of annotation quality for emotion data. In other words, given the subjectivity of the emotion annotation task, group-level *reliability* is used as a replacement for pairwise *agreement*. This, in turn, has important implications for the notion of *ground truth* in emotion annotation.

In a more traditional linguistic annotation setup, the result of a carefully carried out annotation process—in which disagreement between individual annotators can be discussed and resolved—is considered ground truth. In contrast, in emotion annotation, because of its inherent subjectivity, disagreements cannot be resolved. In light of these considerations, I argue that a better way to conceptualize the ground truth behind emotion data would be to consider the aggregated rating when questioning the entire population of all possible raters, i.e., the *expectation* of the rating, a theoretical value that can be estimated but not measured directly.[8] Framing the ground truth in emotion analysis in such a way clarifies why the annotation quality is directly dependent on the number of raters (Buechel and Hahn, 2018a). The reason is that larger studies give more accurate estimates of the population mean. Note that these considerations would make no sense in the context of traditional linguistic annotation, such as in part-of-speech tagging. Although data from one annotation study may be of better quality than that of another one, both studies yield *their own* ground truth because linguistic ground truth is only established through annotation and does not exist independently of it. However, emotion—at least from the Experienced, Perceived, and Evoked viewpoint—is not primarily a linguistic phenomenon (§2.1; §2.3).

Importantly, stipulating that the result of an emotion annotation study is only an *estimate* of the ground truth creates the conceptual opportunity for an NLP system to "outperform" human annotation by providing an even better estimate. In particular, if a system is shown to provide better ground truth estimates for a particular language domain compared to a typically-sized rating study, then this system may replace the majority of manual annotation activities for this domain. Human labor would then be limited to situations where even

[7] This foundation on pairwise comparisons also applies to κ variants that can be used for more than two annotators.

[8] While it may seem strange from the perspective of a linguistic annotation study to consider the underlying population of possible annotators, I emphasize that this is a key principle in the natural and social sciences that also sits at the heart of inferential statistics (Burton, 2000; Sprott, 2000).

better estimates (derived with a larger than usual number of raters) are required, e.g., for test data creation. Evidently, this situation could be described as having achieved "super-human performance" relative to a given number of raters.

In conclusion, it seems fruitful to conceptualize human performance in emotion analysis in terms of how well a rating study of a particular size can estimate the population-wide aggregate. Following this argument, human performance is not an absolute quantity but can only be understood relative to a given number of raters. Future work will have to transform this notion into a rigorous definition of human and super-human performance. Such a definition would allow researchers to go beyond purely descriptive comparisons of model performance vs. rating reliability (Buechel and Hahn, 2017a, 2018c,a; Buechel et al., 2020a) potentially coming to the conclusion that a certain model does indeed achieve super-human performance and that its prediction could in fact replace manual labor. Importantly, however, this would not mean that a certain task has been "solved" since NLP-based estimates of a population mean can always be further improved.

4.4 Towards Human Performance in Emotion Analysis

Finally, I would like to outline two aspects of the human capacity to understand emotion that, to my knowledge, are still missing from today's state-of-the-art systems. Importantly, not only do current methodologies fail to match human performance in these aspects, existing datasets and evaluation methodologies are not even designed to assess them. This underscores the need to advance emotion analysis models along with the infrastructure for validating them in future work.

Firstly, most research activities in emotion analysis to date have their focus on phrases, sentences, or short paragraphs of texts (here jointly referred to as "utterances"), while very little work has been devoted to modeling the overarching emotion of long text documents such as full-length newspaper articles, novels, or business reports (§2.6). The main reason for this can be found in the absence of suitable text-level datasets (§2.4). The creation of such datasets is hampered, not only by the financial burden of having to annotate very long language units, but more importantly by the conceptual problem of what the "overall emotion" of a text is and how it can be operationalized given the short-lived nature of this type of affective state (§2.1). While these considerations may seem like strong reasons to further avoid working in this direction, it must also be noted that *humans* seem to have no problem at all attributing global emotions to long texts, for example speaking of "a sad novel" or a "cheerful play". The question *how* people come up with this sort of judgment requires researchers to address the problem of *text-level compositionality*, i.e., how the overall emotion of a text relates to the emotion of the sentences it contains. While quite a lot of effort has been dedicated to understanding how emotional meaning is composed on the sentence level (e.g., Neviarouskaya et al., 2011; Taboada et al., 2011; Socher et al., 2013), very little work has been directed towards understanding compositional effects in larger units of language. Making progress in this direction would not only allow for new applications of NLP systems but also generate valuable linguistic insight.

Second, so far, computational modeling of emotion has predominantly addressed the prediction of aggregated judgments of *many* human raters (§2.5; §4.3). In contrast, for us in our everyday life, the behavior of *individuals* arguably plays a much larger role. Take the example of electronic text messaging. It is relatively easy for us humans to tell whether someone very close to us, say a romantic partner, is upset, whereas with someone who we do not share such a close connection with, such as a new colleague, it is much more difficult. Apparently, making such *person-level* predictions is an essential skill to us. However, studies expanding on these considerations have been very rare so far (Socher et al., 2011; Gambino and Calvo, 2018; Davani et al., 2021; Troiano et al., 2021a). Perhaps the most obvious way to capture inter-personal differences in emotion understanding is to train multiple models, each one exclusively on data of a single rater. However, this approach of building personalized prediction models seems likely to raise ethical problems. An alternative way of capturing interpersonal variation, without resorting to personalized data and models, would be to broaden the perspective of modeling studies to think of emotion gold labels, not as a single aggregated rating, but rather as a rating distribution in the context of label distribution learning (Geng, 2016). That is, while previous work has mostly only addressed the *mean* of a rating distribution (in the case of numerical labels), treating it as the one and only gold label, one may get a more complete picture of emotional responses by taking into account other properties of the rating distribution as well. In the case of emotion regression problems, these may include other statistical measures, such as the median, standard deviation, interquartile range, or even the percentage of responses each point of the rating scale receives. Either of the two approaches, rater-specific prediction models or emotion distribution learning, would allow to adequately address situations in which the assumption of a single, homogeneous population-level ground truth is fundamentally flawed, such as in the example of “Italy defeats France in World Cup Final” given in §2.3.

5 Conclusion

This thesis compiles seven studies on NLP-based emotion analysis using fine-grained label formats. At the center of their shared research interest is the observation that the recent trend towards more information-rich annotation schemes is, in a sense, a double-edged sword. On the one hand, more expressive label formats lead to NLP systems that are empirically adequate, capture a wider range of emotional meaning facets, and are thus more useful in a larger variety of downstream applications. On the other hand, the proliferation of competing label formats in the past years has led to a situation where existing gold data is spread thin, not only across different domains, i.e., natural languages, their registers, and linguistic levels, but also different, incompatible annotation schemes. Thus, data re-usability and comparability between NLP systems are compromised. While the total amount of emotion gold data is considerable, for any chosen domain and label format often very little data is available. This problem is only aggravated by the high annotation costs associated with such complex label formats and the distinction between different viewpoints, e.g., Evoked vs. Perceived emotion.

The submitted studies approach this problem area from a variety of angles. In Buechel and Hahn (2016), we presented one of the first systems for predicting VAD scores from text and introduced an approach to translate its output into basic emotion scores in an effort to make the results comparable to previous work. Buechel and Hahn (2017a) presented EmoBank, an utterance-level dataset with VAD annotations, which is unique in having two kinds of double-annotations. Part of it is also annotated according to basic emotions and the entire corpus is labeled for both the Evoked and Perceived viewpoint. EmoBank thus allows studying the interplay and systematic differences between these choices in corpus design. Buechel et al. (2018) presented the first publicly available dataset of empathy in written language, a specific emotional nuance that had been rarely studied in NLP so far despite its importance for human-machine interaction. Our dataset is also unique in that it features a new annotation methodology that allows it to capture robust ratings from the writer (Experienced) emotion viewpoint. In Buechel and Hahn (2018c), we studied multi-task learning as a way to exploit the information richness of the VAD format, thus training better, more robust models. We found that combining feed-forward networks with high-quality pre-trained embedding models yields word-level predictions comparable to human rating reliability. In Buechel and Hahn (2018a), we proposed a similar model for the task of label-to-label mapping, showing that VAD ratings can be automatically translated to the BE5 format (and vice versa) with level of reliability which, again, is comparable to humans. The underlying mapping models were also found to generalize well across languages. Building on the previous two studies, Buechel et al. (2020a) presented a methodology for generating large, multi-format emotion lexicons in a crosslingual fashion, imposing only relatively mild data requirements. Using this method, we created and released lexicons for 91 languages, each one of them containing hundreds of thousands of entries. Finally, in Buechel et al. (2021), we presented a method that unifies a large part of the previous studies by learning a latent representation of emotion that is agnostic towards label formats, languages, linguistic levels, and model architectures. This method thus allows us to embed the emotion of language items from very heterogeneous contexts into a common, distributional space that serves as an "interlingua for emotion", thus offering a solution to the comparability problem of emotion analysis.

The presented studies have diverse applications in both industry and academia. The datasets presented in Buechel and Hahn (2017a) and Buechel et al. (2018) help developing emotion detection systems that are useful in areas such as economic forecasting, customer service agents, or crisis management (§2.2, §2.4, and §2.6). Lexicons built in Buechel and Hahn (2018c), Buechel and Hahn (2018a) and Buechel et al. (2020a) can be used, e.g., as supplemental input to utterance- or text-level models, for stimulus selection in psychological experiments, and for writing assistance applications (§4.1). The label mapping approach studied in Buechel and Hahn (2016), Buechel and Hahn (2017a), and Buechel and Hahn (2018a) helps to augment the emotional richness of existing datasets, which in turn makes the resulting predictive models more informative and by extension more beneficial for downstream applications (§2.2; §2.8). Finally, the label- and domain-agnostic emotion embeddings introduced in Buechel et al. (2021) may not only serve as a new backbone technology for the field of emotion analysis but also enable a wide range of psychological, linguistic, and cultural follow-up studies by allowing to directly compare emotion ratings from various settings (§2.8).

Still, ample opportunity for future work remains. Perhaps most obviously, the above line of work on generalized emotion embeddings should be extended to cover other modalities besides written language, such as audio and images, possibly even bio-signals. Furthermore, as discussed in §4.3, a more rigorous theoretical foundation of "human performance" and "super-human performance" is dearly needed to better assess the progress made so far. Finally, §4.4 has outlined two directions in which the capabilities in emotion understanding of NLP systems could be extended to better match those of humans: addressing text-level compositionality and attending more closely to interpersonal variance in emotion expression and elicitation.

6 Bibliography

Following the ACL bibliography style, the PDF version of this thesis has hyperlinks to electronic copies of the following references embedded within the document. Links are based on Digital Object Identifiers (DOIs) whenever possible and regular URLs otherwise.

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 506–515.

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

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech and Language*, 50(C):40–61.

Liz Allen, Alison O'Connell, and Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? How the contributor role taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1):71–74.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, pages 196–205.

Silvio Amir, Ramon F. Astudillo, Wang Ling, Bruno Martins, Mario J. Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale Twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 613–618.

Eduardo B. Andrade and Dan Ariely. 2009. The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes*, 109(1):1–8.

Iris Bakker, Theo van der Voordt, Peter Vink, and Jan de Boon. 2014. Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. *Current Psychology*, 33(3):405–421.

Alexandra Balahur, Jesus M. Hermida, and Andres Montoyo. 2012. Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1):88–101.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493.

C. Daniel Batson, Jim Fultz, and Patricia A. Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of Personality*, 55(1):19–39.

Eduardo Bericat. 2016. The sociology of emotions: Four decades of progress. *Current Sociology*, 64(3):491–513.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Margaret M. Bradley and Peter J. Lang. 2007. Affective norms for English text (ANET): Affective ratings of text and instruction manual. Technical Report D-1, The Center for Research in Psychophysiology, University of Florida.

Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. Discrete emotion norms for nouns: Berlin affective word list (DENN–BAAWL). *Behavior Research Methods*, 43:441.

Joost Broekens. 2012. In defense of dominance: PAD usage in computational representations of affect. *International Journal of Synthetic Emotions*, 3(1):33–42.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122.

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Sven Buechel and Udo Hahn. 2017b. A flexible mapping scheme for discrete and dimensional emotion representations: Evidence from textual stimuli. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 180–185.

Sven Buechel and Udo Hahn. 2017c. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904.

Sven Buechel and Udo Hahn. 2018b. Representation mapping: A novel approach to generate high-quality multi-lingual emotion lexicons. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 184–191.

Sven Buechel and Udo Hahn. 2018c. Word emotion induction for multiple languages as a deep multi-task learning problem. 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 1907–1918.

Sven Buechel, Udo Hahn, Jan Goldenstein, Sebastian G. M. Händschke, and Peter Walgenbach. 2016a. Do enterprises have emotions? In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 147–153.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016b. Feelings from the past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities*, pages 54–61.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2017. The course of emotion in three centuries of German text—A methodological framework. In *Digitial Humanities 2017. Conference Abstracts*, pages 176–179.

Sven Buechel, Simon Junker, Thore Schlaak, Claus Michelsen, and Udo Hahn. 2019. A time series analysis of emotional loading in central bank statements. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 16–21.

Sven Buechel, Luise Modersohn, and Udo Hahn. 2021. Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020a. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217.

Sven Buechel, João Sedoc, H. Andrew Schwartz, and Lyle Ungar. 2020b. Learning emotion from 100 observations: Unexpected robustness of deep learning under strong data limitations. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 129–139.

Dawn Burton. 2000. *Research Training for Social Scientists*. Sage.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days*, pages 37–43.

Jean Carletta. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Edward G. Carmines and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. Sage.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

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

Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2158–2168.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Magali Clobert, Tamara L Sims, Jiah Yoo, Yuri Miyamoto, Hazel R Markus, Mayumi Karasawa, and Cynthia S Levine. 2020. Feeling excited or taking a bath: Do distinct pathways underlie the positive affect–health link in the US and Japan? *Emotion*, 20(2):164–178.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 28–34.

Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 2016. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153.

Charles Darwin. 1872/1998. *The expression of the emotions in man and animals*. Oxford University Press. Orig. published 1872.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv:2110.05719*.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2022. Joint emotion label space modeling for affect lexica. *Computer Speech & Language*, 71:101257.

Luna De Bruyne, Orphee De Clercq, and Veronique Hoste. 2020. An emotional mess! Deciding on a framework for building a Dutch emotion-annotated corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1643–1651.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stephane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Proceedings of the 2nd Grand Challenge and Workshop on Multimodal Language*, pages 1–7.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Lingjia Deng and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189.

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

David DeSteno, James J. Gross, and Laura Kubzansky. 2013. Affective science and health: The importance of emotion and emotion regulation. *Health Psychology*, 32(5):474–486.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Catharine Evers, Alexandra Dingemans, Astrid F. Junghans, and Anja Boevé. 2018. Feeling bad or feeling good, does emotion affect your consumption of food? A meta-analysis of the experimental evidence. *Neuroscience and Biobehavioral Reviews*, 92:195–208.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Itsuki Fujisaki, Hidehito Honda, and Kazuhiro Ueda. 2017. On an effective and efficient method for exploiting "wisdom of crowds in one mind". In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 2043–2048.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the Supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91.

Omar Juárez Gambino and Hiram Calvo. 2018. Modeling distribution of emotional reactions in social media using a multi-target strategy. *Journal of Intelligent & Fuzzy Systems*, 34(5):2837–2847.

Suyu Ge, Lu Cheng, and Huan Liu. 2021. Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*, pages 496–506.

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Matej Gjurkovic and J. Snajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

Sunita Goel and Ozlem Uzuner. 2016. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079.

Wan-Jun Guo, Hui-Yao Wang, Wei Deng, Ming-Jin Huang, Zai-Quan Dong, Yang Liu, Shan-Xia Luo, Jian-Ying Yu, Xia Huang, Yue-Zhu Chen, Chun-Tao Shen, Tian-Rui Ren, Wen Wang, Xin Sun, Xiao-Xi Zeng, Lei Chen, Wei-Hong Kuang, Chang-Jian Qiu, Jin-Ping Song, Da-Jiang Li, Yong Zeng, Nan-Sheng Cheng, Wei-Min Li, Wei Zhang, Lan Zhang, and Tao Li. 2019. Effects of anxiety and depression and early detection and management of emotional distress on length of stay in hospital in non-psychiatric inpatients in China: A hospital-based cohort study. *The Lancet*, 394:S83.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1652–1663.

Petr Hajek, Vladimir Olej, and Renata Myskova. 2014. Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20(4):721–738.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

Sebastian G.M. Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. 2018. A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31.

Rachel Harsley, Bhavesh Gupta, Barbara Di Eugenio, and Huayi Li. 2016. Hit songs' sentiments harness public mood & predict stock market. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 17–25.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

Johannes Hellrich. 2018. *Word Embeddings: Reliability & Semantic Change*. IOS Press.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JeSemE: Interleaving semantics and emotions in a Web service for the exploration of language change phenomena. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 10–14.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2019. Modeling word emotion in historical language: Quantity beats supposed stability in seed word selection. In *Proceedings of the Third Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11.

Johannes Hellrich, Stefan Schulz, Sven Buechel, and Udo Hahn. 2015. JUFit: A configurable rule engine for filtering and generating new multilingual UMLS terms. In *AMIA Annual Symposium Proceedings 2015*, pages 604–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Arlie Russell Hochschild. 1983. *The Managed Heart: Commercialization of Human Feeling*. University of California Press.

Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht-Ecklundt, Harald C. Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3316–3320.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138.

Markus J. Hofmann, Lars Kuchinke, Sascha Tamm, Melissa L.-H. Võ, and Arthur M. Jacobs. 2009. Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience*, 9(4):389–397.

Geoff Hollis. 2018. Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, 50:711—729.

Geoff Hollis and Chris Westbury. 2018. When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, 50:115–133.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Tianqiang Hu, Dajun Zhang, Jinliang Wang, Ritesh Mistry, Guangming Ran, and Xinqiang Wang. 2014. Relation between emotion regulation and mental health: A meta-analysis review. *Psychological Reports*, 114(2):341–362.

Qianjia Huang, Diana Inkpen, Jianhong Zhang, and David Van Bruwaene. 2018. Cyberbullying intervention based on convolutional neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, pages 42–51.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, pages 216–225.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The manually annotated sub-corpus of American English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2455–2461.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73.

Nancy Ide and Keith Suderman. 2004. The American national corpus first release. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1681–1684.

Kamil K. Imbir. 2016. Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, 7:1081.

Carroll E. Izard. 1971. *The Face of Emotion*. Appleton-Century-Crofts.

Daniel Jurafsky and James H. Martin. 2021. Lexicons for sentiment, affect, and connotation. In *Speech and Language Processing*. Draft of September 21, 2021.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665.

Philipp Kanske and Sonja A. Kotz. 2010. Leipzig affective norms for German: A reliability study. *Behavior Research Methods*, 42(4):987–991.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. SWAT-MP: The SemEval-2007 systems for task 5 and task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 308–313.

Sopan Khosla, Niyati Chhaya, and Kushal Chawla. 2018. Aff2Vec : Affect–enriched distributional word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2204–2218.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? Annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

Evgeny Kim and Roman Klinger. 2019a. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64.

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

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26.

Svetlana Kiritchenko and Saif Mohammad. 2017. 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.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. 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.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 437–442.

Manfred Klenner. 2016. A model for multi-perspective opinion inferences. In *Proceedings of IJCAI 2016 Workshop Natural Language Meets Journalism*, pages 6–11.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy. A data set and results for classifying emotions across languages. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75.

Agnieszka Landowska. 2018. Towards new mappings between emotion representation models. *Applied Sciences*, 8(2):274.

Peter J. Lang. 1980. Behavioral treatment and bio-behavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, and T. A. Williams, editors, *Technology in Mental Health Care Delivery Systems*, pages 119–137. Ablex, Norwood/NJ.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2009. Cause event representations for happiness and surprise. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 297–306.

Nicolas Leveau, Sandra Jhean-Larose, Guy Denhière, and Ba-Linh Nguyen. 2012. Validating an interlingual metanorm for emotional analysis of texts. *Behavior Research Methods*, 44(4):1007–1014.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Shoushan Li, Jian Xu, Dong Zhang, and Guodong Zhou. 2016. Two-view label propagation to semi-supervised reader emotion classification. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2647–2655.

Ying Li, Tomas Engelthaler, Cynthia S. Q. Siew, and Thomas T. Hills. 2019. The Macroscope: A tool for examining the historical structure of language. *Behavior Research Methods*, 51(4):1864–1877.

Jasy Suet Yan Liew, Howard R. Turtle, and Elizabeth D. Liddy. 2016. EmoTweet-28: A fine-grained emotion corpus for sentiment analysis. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1149–1156.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6293–6298.

Huanhuan Liu, Shoushan Li, Guodong Zhou, Chu-Ren Huang, and Peifeng Li. 2013. Joint modeling of news reader's and comment writer's emotions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 511–515.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 125–132.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution — A case study to get around IPRs and privacy constraints featuring the German JSynCC corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 1259–1266.

Jordan J. Louviere, Terry N. Flynn, and A.A.J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 259–265.

Asifa Majid. 2012. Current emotion research in the language sciences. *Emotion Review*, 4(4):432–443.

Scott W. McQuiggan and James C. Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies*, 65(4):348–360.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03*, pages 139–144.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Gilad Mishne. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, pages 321–327.

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

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 31–41.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Saif Mohammad and Tony Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the Second Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 70–79.

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

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

Catherine Monnier and Arielle Syssau. 2008. Semantic contribution to verbal short-term memory: Are pleasant words easier to remember than neutral words in serial recall and serial recognition? *Memory & Cognition*, 36(1):35–42.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46(3):887–903.

Maria Moritz, Johannes Hellrich, and Sven Büchel. 2018a. A method for human-interpretable paraphrasticality prediction. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 113–118.

Maria Moritz, Johannes Hellrich, and Sven Buechel. 2018b. Towards a metric for paraphrastic modification. In *Digital Humanities 2018. Book of Abstracts*, pages 457–459.

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

Eric T. Nalisnick and Henry S. Baird. 2013. Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. Affect Analysis Model: Novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95–135.

John B. Nezlek, Konstantinos Kafetsios, and C. Veronica Smith. 2008. Emotions in everyday social encounters: Correspondence between culture and self-construal. *Journal of Cross-Cultural Psychology*, 39(4):366–372.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the 9th Joint Conference on Lexical and Computational Semantics*, pages 58–70.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2013. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 4(1):106–115.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. pages 4367–4380.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press.

Ana P. Pinheiro, Marcelo Dias, João Pedrosa, and Ana P. Soares. 2017. Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*, 49(2):698–716.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15.

Ashequl Qadir and Ellen Riloff. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1209.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Arun Ramachandran and Gerard de Melo. 2020. Cross-lingual emotion lexicon induction using representation alignment in low-resource settings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5879–5890.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

Kevin Reschke and Pranav Anand. 2011. Extracting contextual evaluativity. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 370–374.

Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL–R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777.

Masoud Rouhizadeh, Kokil Jaidka, Laura Smith, H. Andrew Schwartz, Anneke Buffone, and Lyle Ungar. 2018. Identifying locus of control in social media language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1152.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Alan G. Sanfey, James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2003. The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626):1755–1758.

Klaus R. Scherer. 2000. Psychological models of emotion. In Joan C. Borod, editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for the universality and cultural variation of differential emotion response pattering. *Journal of Personality and Social Psychology*, 66(2):310–328.

Axel Schulz, Tung Dang Thanh, Heiko Paulheim, and Immanuel Schweizer. 2013. A fine-grained sentiment analysis approach for detecting crisis related microposts. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*, pages 846–851.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1664–1673.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 564–571.

Shuju Shi, Yinglun Sun, Jose Zavala, Jeffrey Moore, and Roxana Girju. 2021. Modeling clinical empathy in narrative essays. In *Proceedings of the 2021 IEEE 15th International Conference on Semantic Computing*, pages 215–220.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. 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.

Swapna Somasundaran, Xianyang Chen, and Michael Flor. 2020. Emotion arcs of student narratives. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 97–107.

D.A. Sprott. 2000. *Statistical Inference in Science*. Springer.

Hans Stadthagen-González, Constance Imbault, Miguel A. Pérez-Sánchez, and Marc Brysbært. 2017. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49:111–123.

Jacopo Staiano and Marco Guerini. 2014. Depeche Mood: A lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433.

Jan Stets and Jonathan H. Turner. 2006. *Handbook of the Sociology of Emotions*. Springer.

Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James. 2007. Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: An affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 1083–1086.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28*, pages 1–9.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Shabnam Tafreshi. 2021. *Cross-Genre, Cross-Lingual, and Low-Resource Emotion Classification*. ProQuest.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104.

Shabnam Tafreshi and Mona Diab. 2018. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565.

Yi-jie Tang and Hsin-Hsi Chen. 2012. Mining sentiment words from microblogs for predicting writer-reader emotion transition. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1226–1229.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021a. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49.

Enrica Troiano, Aswathy Velutharambath, and Roman Klinger. 2021b. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *arXiv:2110.15871*.

Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2Vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2019. Incorporating textual information on user behavior for personality prediction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 177–182.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2009. Writer meets reader: Emotion analysis of social media from both the writer's and reader's perspectives. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 287–290.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xue-jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 788–793.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 165–176.

Feng Zhou, Shu Kong, Charless C. Fowlkes, Tao Chen, and Baiying Lei. 2020. Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*, 392:38–49.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 609–626.

Part II

Articles Submitted for Examination

Part II of this thesis includes the submitted articles as published in their respective venues. Each chapter is dedicated to one of the papers and starts by giving its full bibliographical reference as well as a statement about the individual contribution of the respective authors. The way in which author contributions are described is inspired by the Contributor Roles Taxonomy (CRediT; Allen et al., 2019). Yet, roles were added or removed as I deemed fitting for NLP research. Table 6.1 summarizes my contribution to the submitted articles.

Role	**Ch.7**	**Ch.8**	**Ch.9**	**Ch.10**	**Ch.11**	**Ch.12**	**Ch.13**
Conception	✓	✓	✓	✓	✓	✓	✓
Supervision	𝒪	𝒪	𝒪	𝒪	𝒪	✓	𝒪
Project Management	𝒪	𝒪	𝒪	𝒪	𝒪	𝒪	𝒪
Data Acquisition		✓	✓				
Methodology Development	✓	✓	𝒪	✓	✓	✓	✓
Model Development	✓	✓	✓	✓	✓	✓	✓
Software Development	✓	✓	✓	✓	✓	𝒪	✓
Experimental Design	✓	✓	✓	✓	✓	✓	✓
Execution of Experiments	✓	✓	✓	✓	✓	𝒪	✓
Data Analysis and Visualization	✓	✓	✓	✓	✓	✓	✓
Writing	✓	✓	✓	✓	✓	✓	✓

Table 6.1: Overview of my contribution to the submitted articles. Checkmark ("✓") indicates the respective role was (co-)performed by me; circle ("𝒪") means the role has been performed by my co-authors but not me; absence of markers means that the respective role is not applicable for this paper.

7 Emotion Analysis as a Regression Problem

Reference

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122.

Author Contributions

Udo Hahn performed supervision and project administration. I performed methodology, model, and software development, experimental design and execution of experiments, as well as data analysis and visualization. Conception and writing were performed jointly by both authors.

1114

ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1114

Emotion Analysis as a Regression Problem — Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation

Sven Buechel and **Udo Hahn**[1]

Abstract. Emotion analysis (EA) and sentiment analysis are closely related tasks differing in the psychological phenomenon they aim to catch. We address fine-grained models for EA which treat the computation of the emotional status of narrative documents as a regression rather than a classification problem, as performed by coarse-grained approaches. We introduce Ekman's Basic Emotions (BE) and Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model—two major schemes of emotion representation following opposing lines of psychological research, i.e., categorical and dimensional models—and discuss problems when BEs are used in a regression approach. We present the first natural language system thoroughly evaluated for fine-grained emotion analysis using the VAD scheme. Although we only employ simple BOW features, we reach correlation values up until $r = .65$ with human annotations. Furthermore, we show that the prevailing evaluation methodology relying solely on Pearson's correlation coefficient r is deficient which leads us to the introduction of a complementary error-based metric. Due to the lack of comparable (VAD-based) systems, we, finally, introduce a novel method of mapping between VAD and BE emotion representations to create a reasonable basis for comparison. This enables us to evaluate VAD output against human BE judgments and, thus, allows for a more direct comparison with existing BE-based emotion analysis systems. Even with this, admittedly, error-prone transformation step our VAD-based system achieves state-of-the-art performance in three out of six emotion categories, out-performing all existing BE-based systems but one.

1 Introduction

Affective states expressed via written or spoken utterances, as well as non-verbal gestures and mimics in discourse are at the core of any cognitively plausible theory of human communication. From a computational perspective, AI researchers have already started investigating into this field [26], since progress in this area will pave the way to even smarter and more natural computational agents for human-computer interaction, such as avatars or robots .

However, this research area at the intersection of (cognitive) psychology, (computational) linguistics, and artificial intelligence suffers from some confusing uses of terminology [22] which have to be sorted out before we get started. Following Pang and Lee [25] we subsume all work done in this area under the umbrella term *subjectivity analysis*. Its most widespread subtask is sentiment analysis or opinion mining (both terms are used interchangeably). In this work, we address another subtask which has recently become more and more popular, namely *emotion analysis* (EA). From a representational perspective, *sentiment* typically refers to the semantic polarity (the positiveness or negativeness relative to some target entity) of a sentence or a document. While sentiment analysis has usually only loose (or no) ties to models taken from psychology, *emotion* (describing phenomena such as anger, fear, or joy) is often represented in a more complex way making direct use of larger pieces of psychological theory.

There are two main dividing lines in the field of EA. The first one (as discussed, e.g., by Calvo and Kim [10]) relates to the choice of a psychological model. Following *categorical models*, emotional states can be subcategorized into a small set of emotion categories. Ekman's Basic Emotion (BE) model [14] is perhaps the most influential among those categorical approaches. On the other hand, following *dimensional models* an emotional state is described relative to a small number of *emotional dimensions*. Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model [28] is among the most commonly used dimensional approaches.

The second and maybe even more fundamental dividing line (as discussed, e.g., by Strapparava and Mihalcea [34]) relates to the main type of predictive problem one faces here. Most of the previous work on EA is *coarse-grained* in the sense that the task of predicting emotion is phrased as a *classification* problem—the output of a corresponding system represents an emotional value as one or multiple class labels. In contrast, *fine-grained* EA treats the task of recognizing emotions as a *regression* problem so that (most often) a vector of real-valued numbers will be produced as the result of an emotion assessment. Note that the choices regarding these dividing lines are made independently from one another, e.g., also allowing for a coarse-grained analysis using dimensional models [16].

The coarse-grained approach seems to be particularly appropriate for highly opinionated social media texts (such as blogs, chats or tweets) but is less likely to account for more subtle expressions of emotions as, e.g., in literary documents (mainly studied in the emerging field of digital humanities [1, 38]), public and personal health narratives (mainly studied in the field of biomedical and clinical NLP [13, 30]) or socio-economic texts (newspaper, newswire, formal business reporting notes, etc. which are increasingly dealt with in computational social science and economics [3, 15, 9]).

In this paper, we focus exclusively on fine-grained emotion analysis. We, first, provide a critical comparison of the BE and the VAD emotion model, as well as a complete survey of prior systems for fine-grained EA. We then present the first VAD-based system for

[1] Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany, URL: http://www.julielab.de

fine-grained EA. Evaluating its performance revealed systematic deficiencies in the evaluation methodology for such systems which lead us to propose a complementary metric. In an attempt to compare our dimensional system more directly with already existing categorical ones, we developed a novel method for mapping between VAD and BE representation schemes and, given these (imperfect) mappings, we find evidence that our system is still among the best-performing systems for predicting the emotional status of narratives.

2 Related Work

2.1 Dimensional versus Categorical Models

Researchers in NLP and psychology have devised a multitude of different models of emotion which can be roughly subdivided into categorical and dimensional models [29, 10, 33]. In computational studies, categorical models most often employ Ekman's [14] six basic emotions (BE: anger, disgust, fear, joy, sadness and surprise) or a derivative therefrom. According to this psychological theory, all human beings share a common set of cross-culturally universal (basic) emotions so that each emotional state of an individual can be unambiguously classified as one of these. Dimensional approaches, on the other hand, often refer to Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model [28].[2] According to this model, emotional states can be described relative to three fundamental emotional dimensions: Valence (the degree of pleasure or displeasure of an emotion), Arousal (level of mental activity, ranging from low engagement to ecstasy) and Dominance (extent of control felt in a given situation). Accordingly, emotions are characterized on three dimensions, each of which spans an interval of real-valued numbers indicating the strength and orientation on each dimension. Providing a fine-grained representation using the VAD model (a vector of real-valued numbers) is therefore straightforward. For BE models, this is typically accomplished by assigning an agreement score to each of the basic emotions (e.g., in the interval [0,100] as realized in the SemEval-2007 test corpus for the *Affective Text* task [34]).

To further illustrate the relationship between the VAD and the BE model, Figure 1 depicts the position of Ekman's basic emotions within the emotional space spanned by the Valence, Arousal and Dominance axis of the VAD model. The assessments were empirically determined by requesting several subjects to describe the six basic emotions in terms of these three dimensions [28]. For fine-grained approaches, we consider VAD to be superior to BE due to the following considerations:

- As Figure 1 reveals, the basic emotions are unevenly distributed in the VAD space. While half of them (anger, disgust and fear) are marked by high arousal and low valence (and therefore reside in one quarter of the space), none of them exhibits high valence and low arousal specifying an emotion like calmness or content. Thus, trying to detect such emotions using a BE-based system may encounter serious problems. Exactly these kinds of emotions have been shown to be most beneficial for the prediction of stock market prices in previous work [3].
- Although Ekman's six-category system is most commonly used, there is no consensus on a fixed set of basic emotions, neither in psychology [29], nor in AI (cf., e.g., [23] and [32]). Not only does this hamper comparison across systems but also does it force researchers to choose different sets of emotional categories according to the emotions which they think to be most relevant for a given application (instead of using a generic and universal representation scheme). This may lead to study designs (e.g., [13]) using a total of 15 different categories considered to indicate suicidal tendencies, e.g., hopelessness or sorrow.
- It is intuitively clear that BEs are *not* equidistant, e.g., fear is obviously more similar to disgust than it is to joy—an observation also supported by Figure 1. Therefore (unlike vectorial VAD representations), distances between given emotions in fine-grained BE representation cannot be meaningfully calculated assuming a vector space with orthogonal axis. This property seriously limits the possibility for further analysis of emotion distributions (such as clustering) and may pose problems for the use of emotion values as features in machine learning.

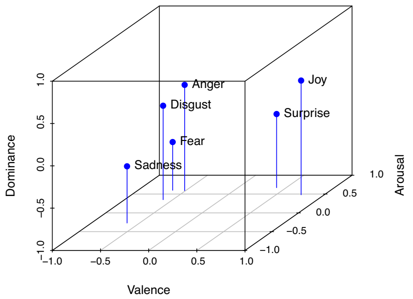


Figure 1. Positions of Ekman's basic emotions within the emotional space spanned by the Valence, Arousal and Dominance axis of the VAD model. Ratings are taken from Russell and Mehrabian [28].

2.2 Computational Resources for Emotion Analysis

In psychology, both models, Ekman's BE as well as Russell and Mehrabian's VAD model, are widely used as standard models [33]. While the VAD model and other dimensional models are commonly preferred in some areas of affective computing [8], NLP researchers, especially those dealing with written documents, almost exclusively subscribe to categorical approaches, most often Ekman's model [10]. As a consequence, these preferences for one model or the other are reflected by the types of resources made available.

Concerning emotion lexicons following the VAD model, the *Affective Norms for English Words* (ANEW) [5] has been most influential in psychological research and was also adapted for many languages other than English [39]. The developers of ANEW asked subjects to rate their feelings on the three VAD dimensions when reading certain words as stimuli. Their responses were encoded using the *Self-Assessment Manikin* (SAM), an icon-style graphical format which consists of three sequences of human-like pictograms, each representing a 9-point scale for Valence, Arousal and Dominance, respectively [4]. The average rating per word was calculated, thus forming its emotional value. The original version of ANEW comprised 1,034 lexical entries. By now, an extended version has been developed amounting to 2,476 words [7].

Bestgen and Vincze [2] extended the original ANEW version by using a bootstrapping method based on Latent Semantic Analysis (LSA) [12]. Their major achievement employing these methods is that they attribute VAD values to formerly unrated words by locating them together with their least distant neighbors whose emotion values are known from the original ANEW resource in a latent semantic

[2] Alternative names for these dimensions include *Pleasure* instead of *Valence* (PAD) as well as *Control* instead *Dominance* (PAC).

space and averaging these values. Re-assessing words already known from ANEW, they compute correlations ($r = 0.71$, 0.56 and 0.60 for Valence, Arousal and Dominance, respectively) between the original and the bootstrapped values. Their lexical resource (BV) incorporates 17,350 entries.

Warriner et al. [39] replicated and extended the original ANEW lexicon in a crowdscourcing campaign using the *Amazon Mechanical Turk* (AMT). Their resource (WKB) contains more than ten times the entries of ANEW (13,915 in total) and excels with particularly high correlations with the original ratings ($r = 0.95$, 0.76 and 0.80 for VAD, respectively). This result is consistent with earlier findings that non-expert ratings for natural language tasks acquired via AMT are, in fact, of good quality (especially when rating emotions) compared to expert ratings [31].

Concerning BE lexicons with fine-grained ratings, Staiano et al. [32] built DEPECHEMOOD (DM), a lexical resource which contains more than 37k entries. They exploit the functionality of the social news network `rappler.com` in which users may report their "mood" when reading a piece of news. DEPECHEMOOD is constructed by multiplying the document-emotion matrix and the document-term matrix of all available mood-rated articles. The latter was computed using either absolute frequency, normalized frequency and TF-IDF scores, thus leading to three versions of the emotion-term matrix.

Another major resource for tackling emotions is WORDNET-AFFECT (WN-A). It contains both, sentiment assessments (positive, negative, neutral and ambiguous) and a hierarchy of various emotion categories [36, 37]. Though not providing continuous ratings for these categories, previous work on fine-grained analysis has largely relied on this resource (cf. Section 2.4).

Corpora carrying VAD annotations are more than rare. To the best of our knowledge, the *Affective Norms for English Text* (ANET) collection [6] is the only available resource and, up until now, has not been used for NLP tasks. With 120 sentences or short texts, e.g., *"You are lying in bed on a Sunday morning"*, it is truly a tiny little corpus. Its VAD annotations were empirically elicited from subjects using SAM (see above). Most recently, another larger resource (FB) carrying at least Valence and Arousal annotations has been generated [27] which comprises 2,895 FACEBOOK posts rated by two annotators.

Corpora annotated with fine-grained emotion categories are rare, as well. To our knowledge, the corpus provided for the *Affective Text* task of SEMEVAL-2007 [34] is the only one, whereas for coarse-grained annotations, there are much more alternatives; cf. [10, 23]. The SEMEVAL corpus (SE7) contains headlines from major newspapers and consists of two subsets, a development set handed out to the competitors (250 headlines) and a final test set (1,000 headlines). The corpus was independently labeled by six annotators according to the BE model so that an agreement score ranging between $[0, 100]$ could be determined for each headline and emotion. Our survey of computational resources is summarized in Table 1.

Two studies [34, 31] report inter-annotator agreement (IAA) measurements for fine-grained BE labeling (see Table 2). Here, IAA is typically measured, first, by calculating Pearson's correlation between each individual annotator and the average annotation of the other annotators (resulting in one correlation value per rater) and then averaging these values [35]. Additionally, Katz et al. [18] provide the agreement of the overlap of their own annotated corpus and SE7. Both are averages of multiple human annotations and are therefore not comparable to IAA values.

Table 1. Resources for emotion detection (lexicons (Lex) and corpora (Corp)) listing the model of emotion they use, the granularity of ratings (Grain), the acquisition methodology (manual (without further specification), asking subjects in a controlled experimental environment (exp), bootstrapping or crowdsourcing (boot or crowd, respectively) and their size in terms of lexical entries (for lexicons) or sentences/documents (for corpora).

	Acronym	Study	Model	Grain	Method	Size
Lex						
	WN-A	[36, 37]	BE	coarse	manual	1,637
	ANEW	[5]	VAD	fine	exp	1,034
	BV	[2]	VAD	fine	boot	17,350
	WKB	[39]	VAD	fine	crowd	13,915
	DM	[32]	BE	fine	crowd[3]	37,771
Corp						
	ANET	[6]	VAD	fine	exp	120
	SE7	[34]	BE	fine	exp	1,250
	FB	[27]	VA	fine	exp	2,895

Table 2. IAA for fine-grained emotion detection measured in r. From the many IAA values reported by Snow et al. [31], we here include their expert *vs.* expert IAA measurements. For comparison, the average is computed only taking anger, fear, joy and sadness into account.

Study	Anger	Disg.	Fear	Joy	Sadness	Surpr.	**Avg.**
[34]	.496	.445	.638	.599	.682	.361	.604
[31]	.459	.583	.711	.596	.645	.464	.603

The IAA presented in the first two studies—ranging between approximately $r = 0.35$ and 0.70—illustrates the hardness of the task.

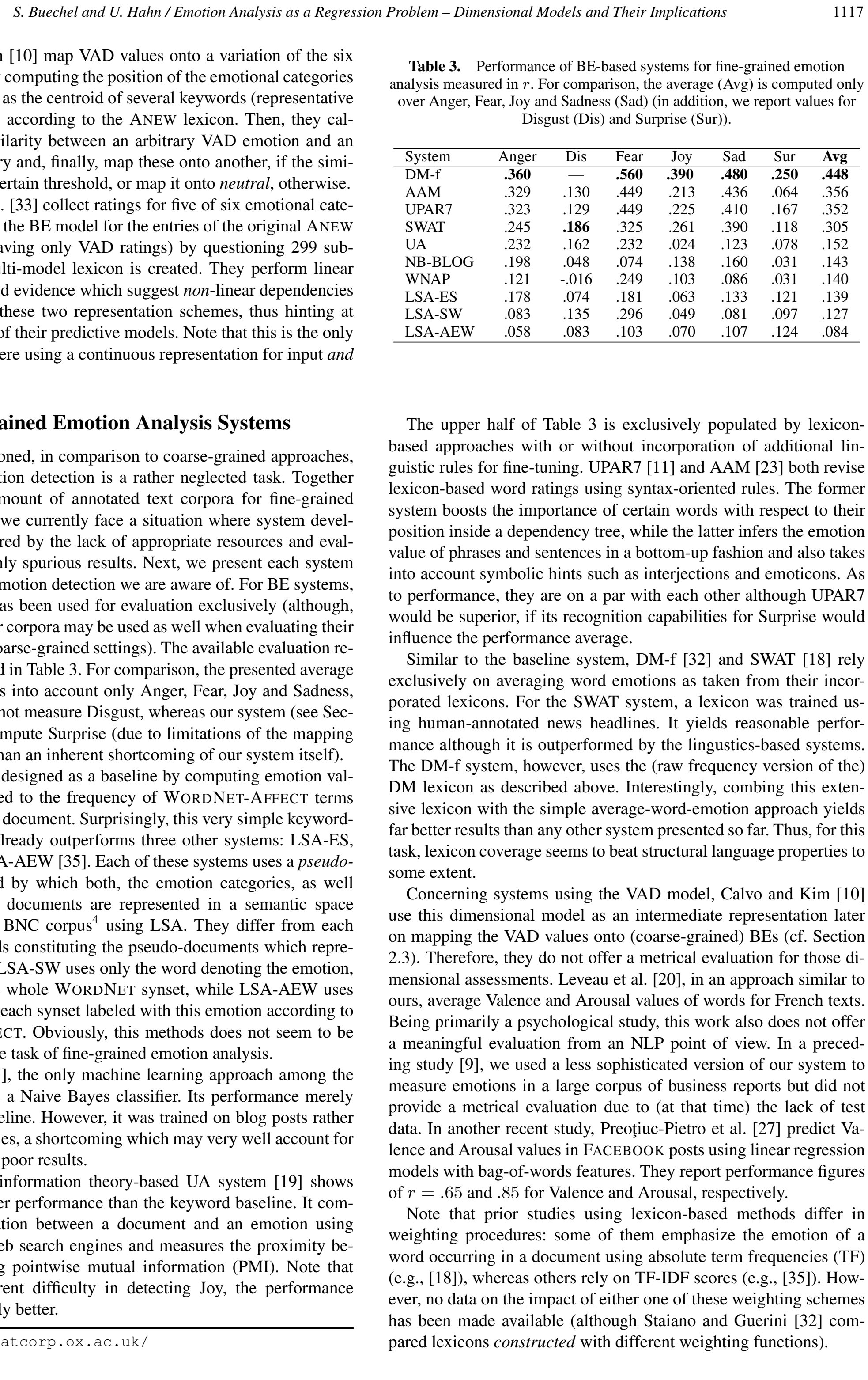
In contrast to BE-based corpora, no IAAs are provided for the VAD-based ANET corpus. However, the average standard deviation between ratings for the same instance amounts to $\text{SD} = 1.45$, 1.85 and 1.87 for Valence, Arousal and Dominance, respectively. The fact that the ratings for the latter two are less consistent than for the former one has been observed in a multitude of studies comparing word ratings, as well as whole lexicons [39, 2]. Preoţiuc-Pietro et al. [27] report an IAA on their FB corpus of $r = .768$ and $.827$ for Valence and Arousal, respectively.

2.3 Mappings between Emotion Models

Only few studies deal with the translation between different emotion schemes. Moreover, most of these activities are only concerned with discrete representations of the BE model (i.e., disregarding continuous agreement scores per category). Having a robust, high-accuracy mapping schema for both representations may help further unify both lines of research (in AI, not limited to NLP, as well as in psychology) [33] and would allow for the interchangeable use of resources developed with respect to one model or the other.

In an early study, Russell et al. [28] presented 300 subjects a list of emotion (or feeling) designating words, including terms referring to the basic emotions, and asked them to assess the designated emotions relative to Valence, Arousal and Dominance. The results can thus be used as a simple yardstick for mapping between basic emotions (in discrete representation) and the VAD model (in dimensional representation) as demonstrated in Figure 1. In a similar, much more recent study, Hoffmann et al. [17] asked 70 subjects to position 22 emotion categories (according to the OCC model [24]) in the VAD space via a user-friendly visual tool. They find high inter-subjective consistency between the assessments although variance was markedly higher for Arousal and Dominance.

[3] Rather than directly using crowdsourced word-emotion ratings, DM was calculated using emotionally crowd-annotated newswire material.



Calvo and Kim [10] map VAD values onto a variation of the six basic emotions by computing the position of the emotional categories in the VAD space as the centroid of several keywords (representative for this category) according to the ANEW lexicon. Then, they calculate cosine similarity between an arbitrary VAD emotion and an emotional category and, finally, map these onto another, if the similarity is above a certain threshold, or map it onto *neutral*, otherwise.

Stevenson et al. [33] collect ratings for five of six emotional categories taken from the BE model for the entries of the original ANEW lexicon (so far having only VAD ratings) by questioning 299 subjects. Thus, a multi-model lexicon is created. They perform linear regression and find evidence which suggest *non*-linear dependencies to hold between these two representation schemes, thus hinting at the insufficiency of their predictive models. Note that this is the only study presented here using a continuous representation for input *and* target variables.

2.4 Fine-Grained Emotion Analysis Systems

As already mentioned, in comparison to coarse-grained approaches, fine-grained emotion detection is a rather neglected task. Together with the small amount of annotated text corpora for fine-grained emotion models, we currently face a situation where system development is hampered by the lack of appropriate resources and evaluations deliver only spurious results. Next, we present each system for fine-grained emotion detection we are aware of. For BE systems, the SE7 corpus has been used for evaluation exclusively (although, additionally, other corpora may be used as well when evaluating their performance in coarse-grained settings). The available evaluation results are presented in Table 3. For comparison, the presented average performance takes into account only Anger, Fear, Joy and Sadness, since DM-f does not measure Disgust, whereas our system (see Section 3) fails to compute Surprise (due to limitations of the mapping functions rather than an inherent shortcoming of our system itself).

WNAP [35] is designed as a baseline by computing emotion values directly related to the frequency of WORDNET-AFFECT terms present in a given document. Surprisingly, this very simple keyword-based approach already outperforms three other systems: LSA-ES, LSA-SW and LSA-AEW [35]. Each of these systems uses a *pseudo-document* method by which both, the emotion categories, as well as the individual documents are represented in a semantic space derived from the BNC corpus[4] using LSA. They differ from each other by the words constituting the pseudo-documents which represent an emotion. LSA-SW uses only the word denoting the emotion, LSA-ES adds the whole WORDNET synset, while LSA-AEW uses each synonym of each synset labeled with this emotion according to WORDNET-AFFECT. Obviously, this methods does not seem to be appropriate for the task of fine-grained emotion analysis.

NB-BLOG [35], the only machine learning approach among the BE systems, uses a Naive Bayes classifier. Its performance merely surpasses the baseline. However, it was trained on blog posts rather than news headlines, a shortcoming which may very well account for a great deal of its poor results.

Similarly, the information theory-based UA system [19] shows only slightly better performance than the keyword baseline. It computes the association between a document and an emotion using statistics from Web search engines and measures the proximity between them using pointwise mutual information (PMI). Note that without its apparent difficulty in detecting Joy, the performance would be markedly better.

Table 3. Performance of BE-based systems for fine-grained emotion analysis measured in r. For comparison, the average (Avg) is computed only over Anger, Fear, Joy and Sadness (Sad) (in addition, we report values for Disgust (Dis) and Surprise (Sur)).

System	Anger	Dis	Fear	Joy	Sad	Sur	**Avg**
DM-f	**.360**	—	**.560**	**.390**	**.480**	**.250**	**.448**
AAM	.329	.130	.449	.213	.436	.064	.356
UPAR7	.323	.129	.449	.225	.410	.167	.352
SWAT	.245	**.186**	.325	.261	.390	.118	.305
UA	.232	.162	.232	.024	.123	.078	.152
NB-BLOG	.198	.048	.074	.138	.160	.031	.143
WNAP	.121	-.016	.249	.103	.086	.031	.140
LSA-ES	.178	.074	.181	.063	.133	.121	.139
LSA-SW	.083	.135	.296	.049	.081	.097	.127
LSA-AEW	.058	.083	.103	.070	.107	.124	.084

The upper half of Table 3 is exclusively populated by lexicon-based approaches with or without incorporation of additional linguistic rules for fine-tuning. UPAR7 [11] and AAM [23] both revise lexicon-based word ratings using syntax-oriented rules. The former system boosts the importance of certain words with respect to their position inside a dependency tree, while the latter infers the emotion value of phrases and sentences in a bottom-up fashion and also takes into account symbolic hints such as interjections and emoticons. As to performance, they are on a par with each other although UPAR7 would be superior, if its recognition capabilities for Surprise would influence the performance average.

Similar to the baseline system, DM-f [32] and SWAT [18] rely exclusively on averaging word emotions as taken from their incorporated lexicons. For the SWAT system, a lexicon was trained using human-annotated news headlines. It yields reasonable performance although it is outperformed by the lingustics-based systems. The DM-f system, however, uses the (raw frequency version of the) DM lexicon as described above. Interestingly, combing this extensive lexicon with the simple average-word-emotion approach yields far better results than any other system presented so far. Thus, for this task, lexicon coverage seems to beat structural language properties to some extent.

Concerning systems using the VAD model, Calvo and Kim [10] use this dimensional model as an intermediate representation later on mapping the VAD values onto (coarse-grained) BEs (cf. Section 2.3). Therefore, they do not offer a metrical evaluation for those dimensional assessments. Leveau et al. [20], in an approach similar to ours, average Valence and Arousal values of words for French texts. Being primarily a psychological study, this work also does not offer a meaningful evaluation from an NLP point of view. In a preceding study [9], we used a less sophisticated version of our system to measure emotions in a large corpus of business reports but did not provide a metrical evaluation due to (at that time) the lack of test data. In another recent study, Preoţiuc-Pietro et al. [27] predict Valence and Arousal values in FACEBOOK posts using linear regression models with bag-of-words features. They report performance figures of $r = .65$ and $.85$ for Valence and Arousal, respectively.

Note that prior studies using lexicon-based methods differ in weighting procedures: some of them emphasize the emotion of a word occurring in a document using absolute term frequencies (TF) (e.g., [18]), whereas others rely on TF-IDF scores (e.g., [35]). However, no data on the impact of either one of these weighting schemes has been made available (although Staiano and Guerini [32] compared lexicons *constructed* with different weighting functions).

[4] http://www.natcorp.ox.ac.uk/

3 Experiments Using Dimensional Models

We start in Section 3.1 by defining a metrical criterion which guides the emotion analysis for JEMAS (Jena Emotion Analysis System),[5] our bag-of-words (BOW) engine (similar to [32] and [18]) employing the VAD model. In Section 3.2, we then evaluate JEMAS using different configurations and discuss implications of these experiments concerning metrical evaluation in Section 3.3.

3.1 Simple Metrics for Emotion Analysis

We distinguish two basic data containers. First, the set of documents (1) where λ denotes some weighting function for terms and $t_{i,j}$ denotes some morphologically normalized non-stop word term in the document-term vector for document d_i, $j = 1, ..., n$; n being the total size of the normalized vocabulary in DOC, so that $\lambda_{t_{i,j}}$ denotes the numerical weight of the *j-th* term from document d_i. Second, the VAD lexicon (2) where each emotion-sensitive lemma lex_l contained in VAD is associated with its corresponding VAD triple $\langle v_l, a_l, d_l \rangle \in \mathbb{R}^3$; each of the three components ranges in the normalized interval $[-4, 4]$, with $l = 1, .., t$; t enumerating the total size of the lexicon.

$$DOC := \{d_i = (\lambda_{t_{i,1}}, ..., \lambda_{t_{i,n}})\} \tag{1}$$

$$VAD := \{vad_l = (lex_l, \langle v_l, a_l, d_l \rangle)\} \tag{2}$$

We may then define the *Emotion Value* of each document d_i (using the projection $\pi_1(VAD) := \{lex \mid (lex, \langle v, a, d \rangle) \in VAD\}$ and the string equality function SEQ):

$$EV_{d_i} := \frac{\sum_{k=1 \,\wedge\, \exists lex_q \in \pi_1(VAD):\ SEQ(lex_q, t_{i,k})}^{n} \lambda_{t_{i,k}} \times \langle v_q, a_q, d_q \rangle}{\sum_{k=1 \,\wedge\, \exists lex_q \in \pi_1(VAD):\ SEQ(lex_q, t_{i,k})}^{n} \lambda_{t_{i,k}}} \tag{3}$$

The general purpose of the term weighting functions λ is to capture the importance a given term, $t_{i,j}$, has for a document d_i. For the following experiments, we specify two such weighting functions (although any other term weighting function for document-term vectors can be employed in this framework). The first weighting function we use, λ_1, is the absolute frequency of a term in a document, $\text{TF}_{i,j}$, that is simply the count how often term $t_{i,j}$ occurs in document d_i:

$$\lambda_1 := \text{TF}_{i,j} \tag{4}$$

Secondly, we use the TF-IDF metric which is the most common weighting scheme in information retrieval [21]. Let $|DOC|$ be the total number of documents in the document collection and let DF_j be the number of documents in which t_j occurs. Hence, our second weighting scheme, λ_2, is defined by the TF-IDF weight of term t_j within the entire document collection:

$$\lambda_2 := \text{TF}_{i,j} \times log \frac{|DOC|}{\text{DF}_j} \tag{5}$$

3.2 Evaluation of the JEMAS Emotion Analyzer

This formal sketch is flexible enough to process documents of arbitrary length, i.e. ranging from a single word to hundreds of pages of full text [9]. However, in the following experiment, we use ANET [6] as a test corpus for the JEMAS system. We transform the VAD ratings associated with the 120 short texts into the interval $[-4, 4]$, with '0' as the neutral rating point for each of the three VAD dimensions. Concerning the chosen lexicons, we decided to compare all of the three lexicons introduced in Section 2 incorporating the VAD model of emotion since they vary largely in terms of size and the underlying acquisition methodology, i.e.,

- the extended (2010-) version of ANEW [7] which—although being rather small—was compiled using a controlled experimental environment,
- the BV lexicon [2] assembled via bootstrapping from the original 1999-version of ANEW [5], and
- the WKB lexicon [39] which reproduces and extends the original ANEW by crowdsourcing.

We transform the emotion value of each lexicon entry so that they are balanced in the interval [–4,4] to simplify interpretation (in the original lexicons, they range in the interval [1,9]).

Since no data on the impact of different term-weighting schemes is available (cf. Section 2.4), we generate results for both, TF and TF-IDF schemes, for a total of six configurations of our system (one for each combination of lexicon and weighting function). Table 4 presents the evaluation results (given in Pearson's correlation) for this experiment.

Table 4. Results of the JEMAS system (Pearson's r) relative to the three VAD dimensions. Evaluation was performed against the ANET corpus with all possible combinations of lexicons and weighting functions.

	Valence		Arousal		Dominance		Avg.	
	tf	tfidf	tf	tfidf	tf	tfidf	tf	tfidf
ANEW	0.53	0.56	0.58	0.58	0.43	0.46	0.51	0.53
BV	0.67	0.68	0.49	0.48	0.66	0.65	0.61	0.61
WKB	0.70	0.71	0.63	0.64	0.59	0.59	0.64	0.65

In general, we find the correlation to the human ratings to be between $r = 0.43$ and 0.71 depending on the lexicon, the weighting function and especially the respective emotional dimension. The crowdsourced and high-volume WKB lexicon provides the best average correlation over Valence, Arousal and Dominance. The BV lexicon gets slightly worse performance figures but still mostly exceeds those that can be achieved using ANEW (except for Arousal). Hence, in terms of performance with respect to the lexicons, coverage seems to beat quality to some extent.

These findings can be further connected to the *recognition rate* of our system, i.e., the percentage of content words in a document which can be attributed an emotion value by our system, using one of the three lexicons: we obtain 42%, 95%, and 87% recognition using the ANEW, the BV, and the WKB lexicon, respectively.

The data suggest that the performance boost of BV and WKB over ANEW can be well explained by superior coverage, whereas the coverage gain BV shows in comparison with WKB seems to be more than compensated by WKB's superior quality due to human ratings as opposed to the semi-supervised approach underlying BV. Note that the BV lexicon used the 1999 version of ANEW as seed set for bootstrapping but still yields better results than the 2010 edition (which

[5] JEMAS will be publicly available on our GITHUB site `https://github.com/JULIELab`.

has more than twice the amount of entries), thus demonstrating the validity of Bestgen and Vincze's [2] bootstrapping method.

Concerning the comparison of TF versus TF-IDF weighting functions, our data (see Table 4) hint at a slight advantage when using TF-IDF scores leading to an increased correlation in seven instances while decreasing it in only two (for Arousal and Dominance using BV). Also, average performance increased by one, respectively two percent points for WKB and ANEW while it remains unchanged for BV. A possible explanation for this improvement could be that common words are emotionally rather neutral and rating consistency is rather poor for emotionally neutral words [39]. Therefore, words whose emotion values are less reliable may be attributed less relevance using TF-IDF resulting in an overall gain in performance.

Of course, our results are not directly comparable to the ones from prior evaluation rounds as shown in Table 3 due to different test corpora and models of emotion. However, it should be noted that the correlation our system obtains with human ratings for the ANET corpus (concerning the VAD emotions) widely exceeds the correlation any of the systems revealed when they are evaluated against the SEMEVAL corpus (in relation to Ekman's six basic emotions) and even exceeds human IAA for two different studies (Table 2). This result is even more exciting since our methodology resembles that of those prior systems, especially DM-f [32], which also employs a broad-coverage emotion lexicon and, in essence, averages word emotion values. We carefully interpret this observation as possibly hinting at the superiority of the VAD model (in terms of its suitability for inter-subjective and reliable assessments for humans, as well as for algorithms) compared with the BE model, a stipulation we further elaborate after the discussion of further experiments below.

Comparing our findings to those of Preoţiuc-Pietro et al. [27], it becomes apparent that performance in emotion analysis strongly depends on the specific domain, i.e., they report a performance of only $r = .113$ and .188 for Valence and Arousal, respectively, using the WKB lexicon on their FACEBOOK posts corpus (in contrast to our system performing at $r = .70$ and .65 using a very similar set-up on the ANET corpus) while linear regression models using BOW features perform at $r = .65$ and .85.

Extending the usual evaluation methodology for fine-grained emotion detection, we decided not only to measure the performance of our system with respect to Pearson's correlation but to also take into account root-mean-square error (RMSE) which is commonly used to assess the quality of a regression model. It is computed as the quadratic mean of the errors, i.e., the differences between the values predicted by the model and the values actually observed. Table 5 displays the same data as in Table 4 for RMSE instead of r.

Table 5. Results of the JEMAS system (RMSE) relative to the three VAD dimensions. Evaluation was performed against the ANET corpus with all combinations of lexicons and weighting functions.

	Valence		Arousal		Dominance		Avg.	
	tf	tfidf	tf	tfidf	tf	tfidf	tf	tfidf
ANEW	2.38	2.33	1.80	1.82	1.78	1.75	1.98	1.97
BV	2.42	2.41	2.03	2.04	1.79	1.79	2.08	2.08
WKB	2.26	2.23	2.57	2.56	1.80	1.78	2.21	2.19

The surprising result of applying RMSE for these configurations is that the relative performance of the three lexicons when compared to one another changes completely. While with r WKB outperfomed BV which itself yielded better results than ANEW, using RMSE, the order of the lexicons according to the measured performance figures is actually reversed (note that since RMSE denotes a measure of error, the lower the value the better the performance).

To further investigate this astonishing result, we plotted the data (only TF-based results) in nine scatterplots (see Figure 2) where each row (with three plots each) displays the results for one lexicon and each column depicts the results for one emotional dimension. Accordingly, a data point in a particular plot denotes the predicted value for an instance of ANET (x-axis) in one emotional dimension using one of the three lexicons and its actual value according to the human ratings (y-axis). The red lines designate the regression line (using a linear model) while the green lines (for comparison) denote a perfect agreement (predicted values equal actual values).

Building on these data visualizations, we venture to cautiously explain the opposing result in terms of r and RMSE. As can be seen, the data points scatter loosely around the regression line when using the ANEW lexicon, whereas for BV and WKB they stick considerably closer to it. Since the (vertical) distance of a data point to the regression line is related to the linear relationship between the two data series, this observation visually "explains" that r values are getting higher from the top line to the bottom line of the scatterplots.

Also, it can be seen that the slope of the regression is much steeper when using the BV and WKB lexicon. The slope of the regression line is related to the interval the predicted values are ranging in. As can be observed, x-values ranging in a small interval result in a steeper slope. This means that data points can be positioned closely to the regression line while at the same time (because of its slope) being far away from the green line (denoting a considerable difference between predicted and actual value).

For instance, in the middle column, the actual value of an instance may be, say, 3 so perfect agreement would demand for a predicted value of 3 as well (as marked by the green line). However, for such instances, our system usually predicts (approximately) a value of 0 (as can be seen) resulting in a large *squared* error. At the same time, the data point being close to the regression line contributes to a high Pearson's correlation (r). Also note that the data points for predicting Arousal with the WKB lexicon (bottom center plot) are off-center (a property most probably derived from the lexicon itself [39]) resulting in an even higher squared error.

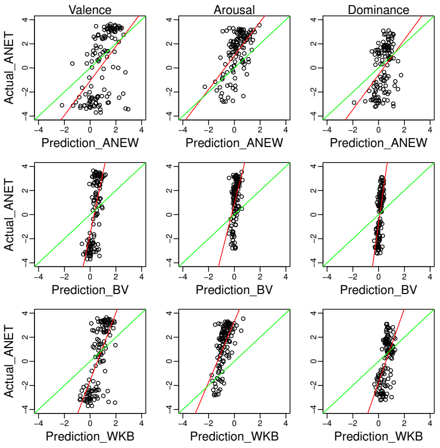


Figure 2. Scatterplotts for a graphical interpretation of the evaluation against the ANET corpus using TF weights. Each data point in each plot designates a pair of a predicted value (x-axis) and the actual value according to ANET (y-axis). The plots are grouped by lexicons used for the evaluation (row-wise) and by emotional dimensions (column-wise).

The steepness of the slopes seems to correlate with the number of entries in the lexicon used to produce the particular data, as well as with the recognition rate (see above). This seems to indicate that the bigger the lexicon, the larger the error caused by this effect could be. A possible explanation for these findings is that most of the words contained in a large emotion lexicon, in contrast to a small one, are on average less emotional (because strongly emotion-bearing lexemes will most likely already be included in a small lexicon, right from the beginning). We conclude that for high performance in terms of Pearson's correlation, the relative differences between the predicted values should be reliable but still the numeric values may differ a lot from the actual values making *any* system unreliable.

The consequences of the above interpretation may, to some extent, be dramatic. Arguably, the prevailing performance measure (Pearson's r) used up until now captures only half of our human intuition of textual emotion, i.e. how the emotion associated with one linguistic unit relates to that of another one—this aspect of a model's predictive power is captured by correlation. It does, however, not capture our ability to perceive the strength and orientation of an emotion with respect to an absolute scale (e.g., neutral arousal vs. highest arousal)—that aspect of a model's predictive power is captured by an error-based metric. While the former may be sufficient for some tasks, it may be irrelevant for others. Therefore, our findings point out that the common evaluation methodology for fine-grained emotion detection is seriously flawed which casts doubt on the validity of prior results (Table 3). Furthermore, since the phenomenon described above was observed using a lexicon-based method (it became more pronounced the larger the coverage of such a resource is), it seems quite likely that, e.g., DM-f, the best-performing system, displays a similar behavior due to the commonalities of the two approaches (ours and theirs). For future work, we therefore suggest to use RMSE as a performance measure complementary to r because taking account of error might be more relevant than the consideration of correlation for many applications and must therefore be addressed during evaluation.

3.3 A Linear Regression-Based Repair Mechanism

In a first attempt to cope with the newly discovered weaknesses of our system, we developed a simple, yet effective repair mechanism to better fit our predictions to the actual data. For each combination of lexicon, weighting function and emotional dimension according to the VAD model, we trained a linear regression model (18, in total) using the originally predicted value of the particular emotion as the only (input) feature. Training was conducted using the Anet corpus. We did not perform cross-validation because these models cannot overfit due to their simplicity. We then post-processed our data from the previous experiment using these models. Table 6 depicts the results of this experiment using RMSE as evaluation yardstick. Note that Pearson's correlation remains unchanged by this procedure.

Table 6. Evaluation result against the Anet corpus after linear regression-based repair (measurements in RMSE).

	Valence		Arousal		Dominance		Avg.	
	tf	tfidf	tf	tfidf	tf	tfidf	tf	tfidf
Anew	2.23	2.18	1.40	1.41	1.72	1.69	1.78	1.76
BV	1.95	1.93	1.50	1.51	1.42	1.44	1.62	1.62
WKB	1.87	1.85	1.33	1.32	1.54	1.53	1.58	1.57

As can be seen, Table 6 resembles Table 5 in many key features, e.g., TF-IDF yields slightly better results than TF (demonstrating the robustness of this method). However, each single RMSE value experienced a pronounced drop of error so that the higher the error was, the more the RMSE decreased, thus rearranging the relative performance figure between the lexicons. After repair, according to RMSE measurements, WKB yields better results than BV which itself is better than Anew. Thus, the order has been reversed in comparison to the data without repair. Furthermore, orderings are now consistent with the ones when using r as performance measure.

The visual interpretation of this method is that instead of predicting the output value of our system, we predict the point on the regression line (displayed in red in Figure 2) associated with it, that is the point above its value on the x-axis. As a result, the new regression line is identical to the line of perfect agreement (displayed in green in Figure 2). We conclude that our method yields satisfactory results despite its simplicity. Yet, the corpus we use for this experiment is quite small (120 instances) so that our method could be less effective when applied to other data sets.

4 Comparison with Categorical Systems

In previous work, we have demonstrated the practical value the VAD data our system produces may have for other areas of research, e.g., emotional portrays of enterprises based on their business and sustainability reports [9]. In contrast, the following section addresses the mapping from VAD to BE representation as a methodological exercise only for the sake of comparison since the number of directly comparable systems is otherwise extremely limited.

4.1 Mapping Emotion Models

As already discussed, being able to reliably convert between different models of emotions, such as the VAD and the BE model, yields many benefits, including better reusability of resources, as well as better means of comparing emotion detection systems using different representation schemes for emotions. Building on the work of Stevenson et al. [33], we here use their complementary BE-based emotional ratings for the Anew lexicon to generate a variety of regression models. We start by transforming Anew's VAD and BE ratings so that the former are balanced in the interval $[-4, 4]$, as we already did with our lexicon, and the latter span the interval $[0, 100]$ so that their interval equals that of the SemEval-2007 test corpus. We used the R Caret package[6] to train linear models, SVMs with a polynomial kernel and kNN models for regression. For either way of the emotion model mapping (VAD → BE, as well as BE → VAD), we trained an independent model for each category or dimension of the emotion representation we map onto, while each category or dimension of the input representation was used as a feature.

For example, when transforming basic emotion to their VAD representation, we trained three independent models each one relying on all of the basic emotions as features. As our models are solely based on the input emotion values (not taking into account other features), they are independent from the type of stimulus eliciting the emotion, e.g. be it a word, a sentence, a text or an image. The performance of these models (obtained using 10-fold cross-validation) is summarized in Tables 7 and 8 using R^2. These values are consistent with the RMSE-based results. Note that, since each table cell represents an independent model, tuning parameter selection may differ across a particular line.

[6] `http://topepo.github.io/caret/index.html`

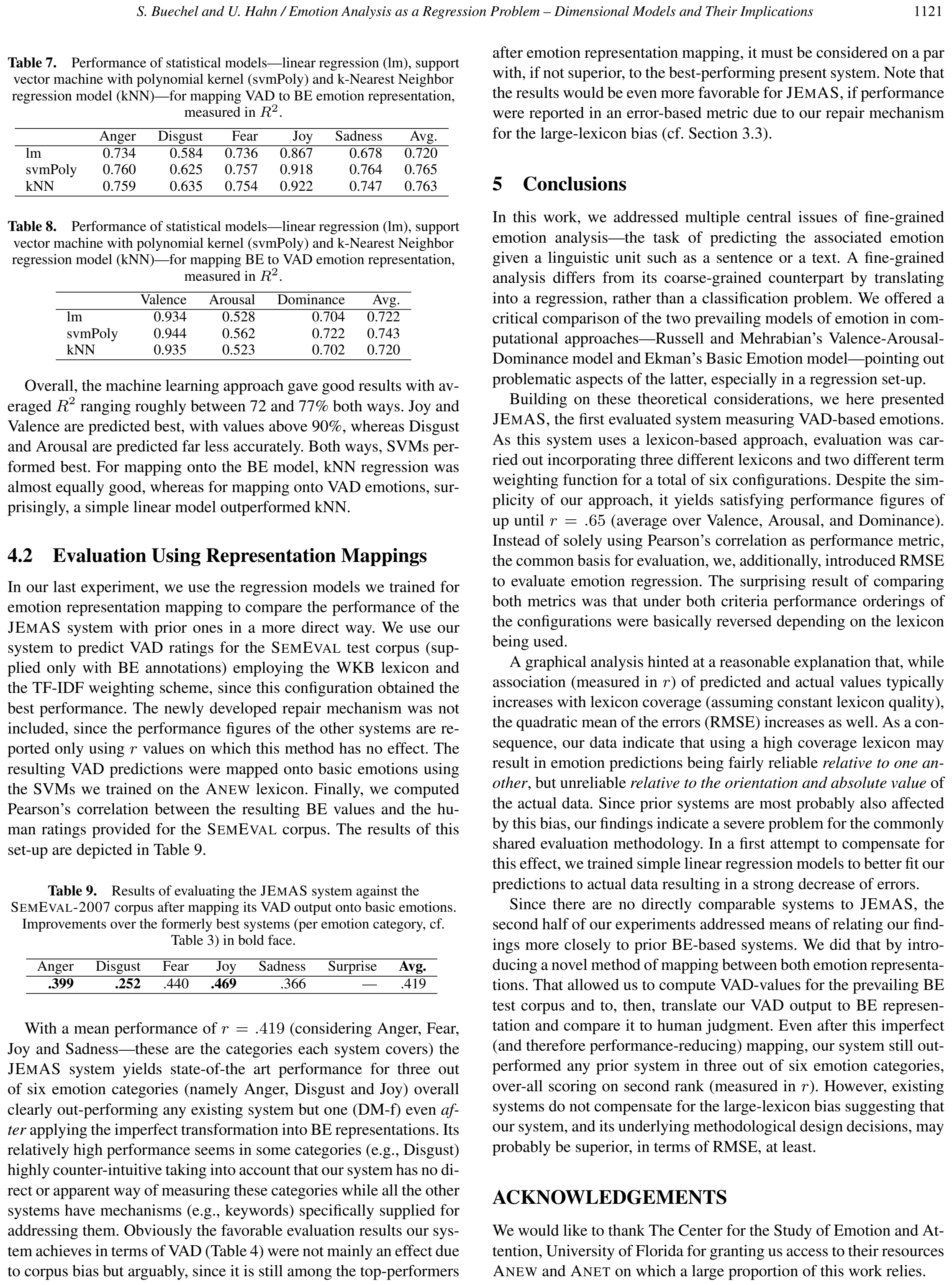


Table 7. Performance of statistical models—linear regression (lm), support vector machine with polynomial kernel (svmPoly) and k-Nearest Neighbor regression model (kNN)—for mapping VAD to BE emotion representation, measured in R^2.

	Anger	Disgust	Fear	Joy	Sadness	Avg.
lm	0.734	0.584	0.736	0.867	0.678	0.720
svmPoly	0.760	0.625	0.757	0.918	0.764	0.765
kNN	0.759	0.635	0.754	0.922	0.747	0.763

Table 8. Performance of statistical models—linear regression (lm), support vector machine with polynomial kernel (svmPoly) and k-Nearest Neighbor regression model (kNN)—for mapping BE to VAD emotion representation, measured in R^2.

	Valence	Arousal	Dominance	Avg.
lm	0.934	0.528	0.704	0.722
svmPoly	0.944	0.562	0.722	0.743
kNN	0.935	0.523	0.702	0.720

Overall, the machine learning approach gave good results with averaged R^2 ranging roughly between 72 and 77% both ways. Joy and Valence are predicted best, with values above 90%, whereas Disgust and Arousal are predicted far less accurately. Both ways, SVMs performed best. For mapping onto the BE model, kNN regression was almost equally good, whereas for mapping onto VAD emotions, surprisingly, a simple linear model outperformed kNN.

4.2 Evaluation Using Representation Mappings

In our last experiment, we use the regression models we trained for emotion representation mapping to compare the performance of the JEMAS system with prior ones in a more direct way. We use our system to predict VAD ratings for the SEMEVAL test corpus (supplied only with BE annotations) employing the WKB lexicon and the TF-IDF weighting scheme, since this configuration obtained the best performance. The newly developed repair mechanism was not included, since the performance figures of the other systems are reported only using r values on which this method has no effect. The resulting VAD predictions were mapped onto basic emotions using the SVMs we trained on the ANEW lexicon. Finally, we computed Pearson's correlation between the resulting BE values and the human ratings provided for the SEMEVAL corpus. The results of this set-up are depicted in Table 9.

Table 9. Results of evaluating the JEMAS system against the SEMEVAL-2007 corpus after mapping its VAD output onto basic emotions. Improvements over the formerly best systems (per emotion category, cf. Table 3) in bold face.

Anger	Disgust	Fear	Joy	Sadness	Surprise	**Avg.**
.399	**.252**	.440	**.469**	.366	—	.419

With a mean performance of $r = .419$ (considering Anger, Fear, Joy and Sadness—these are the categories each system covers) the JEMAS system yields state-of-the art performance for three out of six emotion categories (namely Anger, Disgust and Joy) overall clearly out-performing any existing system but one (DM-f) even *after* applying the imperfect transformation into BE representations. Its relatively high performance seems in some categories (e.g., Disgust) highly counter-intuitive taking into account that our system has no direct or apparent way of measuring these categories while all the other systems have mechanisms (e.g., keywords) specifically supplied for addressing them. Obviously the favorable evaluation results our system achieves in terms of VAD (Table 4) were not mainly an effect due to corpus bias but arguably, since it is still among the top-performers after emotion representation mapping, it must be considered on a par with, if not superior, to the best-performing present system. Note that the results would be even more favorable for JEMAS, if performance were reported in an error-based metric due to our repair mechanism for the large-lexicon bias (cf. Section 3.3).

5 Conclusions

In this work, we addressed multiple central issues of fine-grained emotion analysis—the task of predicting the associated emotion given a linguistic unit such as a sentence or a text. A fine-grained analysis differs from its coarse-grained counterpart by translating into a regression, rather than a classification problem. We offered a critical comparison of the two prevailing models of emotion in computational approaches—Russell and Mehrabian's Valence-Arousal-Dominance model and Ekman's Basic Emotion model—pointing out problematic aspects of the latter, especially in a regression set-up.

Building on these theoretical considerations, we here presented JEMAS, the first evaluated system measuring VAD-based emotions. As this system uses a lexicon-based approach, evaluation was carried out incorporating three different lexicons and two different term weighting function for a total of six configurations. Despite the simplicity of our approach, it yields satisfying performance figures of up until $r = .65$ (average over Valence, Arousal, and Dominance). Instead of solely using Pearson's correlation as performance metric, the common basis for evaluation, we, additionally, introduced RMSE to evaluate emotion regression. The surprising result of comparing both metrics was that under both criteria performance orderings of the configurations were basically reversed depending on the lexicon being used.

A graphical analysis hinted at a reasonable explanation that, while association (measured in r) of predicted and actual values typically increases with lexicon coverage (assuming constant lexicon quality), the quadratic mean of the errors (RMSE) increases as well. As a consequence, our data indicate that using a high coverage lexicon may result in emotion predictions being fairly reliable *relative to one another*, but unreliable *relative to the orientation and absolute value* of the actual data. Since prior systems are most probably also affected by this bias, our findings indicate a severe problem for the commonly shared evaluation methodology. In a first attempt to compensate for this effect, we trained simple linear regression models to better fit our predictions to actual data resulting in a strong decrease of errors.

Since there are no directly comparable systems to JEMAS, the second half of our experiments addressed means of relating our findings more closely to prior BE-based systems. We did that by introducing a novel method of mapping between both emotion representations. That allowed us to compute VAD-values for the prevailing BE test corpus and to, then, translate our VAD output to BE representation and compare it to human judgment. Even after this imperfect (and therefore performance-reducing) mapping, our system still outperformed any prior system in three out of six emotion categories, over-all scoring on second rank (measured in r). However, existing systems do not compensate for the large-lexicon bias suggesting that our system, and its underlying methodological design decisions, may probably be superior, in terms of RMSE, at least.

ACKNOWLEDGEMENTS

We would like to thank The Center for the Study of Emotion and Attention, University of Florida for granting us access to their resources ANEW and ANET on which a large proportion of this work relies.

REFERENCES

[1] Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley, 'The expression of emotions in 20th century books', *PLoS ONE*, **8**(3), e59030, (2013).

[2] Yves Bestgen and Nadja Vincze, 'Checking and bootstrapping lexical norms by means of word similarity indexes', *Behavior Research Methods*, **44**(4), 998–1006, (2012).

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng, 'Twitter mood predicts the stock market', *Journal of Computational Science*, **2**(1), 1–8, (2011).

[4] Margaret M. Bradley and Peter J. Lang, 'Measuring emotion: The self-assessment manikin and the semantic differential', *Journal of Behavior Therapy and Experimental Psychiatry*, **25**(1), 49–59, (1994).

[5] Margaret M. Bradley and Peter J. Lang, 'Affective norms for English words (Anew): Stimuli, instruction manual and affective ratings', Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, (1999).

[6] Margaret M. Bradley and Peter J. Lang, 'Affective norms for English text (Anet): Affective ratings of text and instruction manual', Technical Report D-1, University of Florida, Gainesville, FL, (2007).

[7] Margaret M. Bradley and Peter J. Lang, 'Affective Norms for English Words (Anew): Stimuli, Instruction Manual and Affective Ratings', Technical Report C-2, University of Florida, Gainesville, FL, (2010).

[8] Joost Broekens, 'In defense of dominance: PAD usage in computational representations of affect', *International Journal of Synthetic Emotions*, **3**(1), 33–42, (2012).

[9] Sven Buechel, Udo Hahn, Jan Goldenstein, Sebastian G. M. Händschke, and Peter Walgenbach, 'Do enterprises have emotions?', in *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, pp. 147–153, (2016).

[10] Rafael A. Calvo and Sunghwan Mac Kim, 'Emotions in text: Dimensional and categorical models', *Computational Intelligence*, **29**(3), 527–543, (2013).

[11] François-Régis Chaumartin, 'UPAR7: A knowledge-based system for headline sentiment tagging', in SemEval-2007 *— Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 422–425, (2007).

[12] Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman, 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, **41**(6), 391–407, (1990).

[13] Bart Desmet and Véronique Hoste, 'Emotion detection in suicide notes', *Expert Systems with Applications*, **40**(16), 6351–6358, (2013).

[14] Paul Ekman, 'An argument for basic emotions', *Cognition & Emotion*, **6**(3-4), 169–200, (1992).

[15] Petr Hájek, Vladimír Olej, and Renáta Myšková, 'Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making', *Technological and Economic Development of Economy*, **20**(4), 721–738, (2014).

[16] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu, 'EmoTex: Detecting emotions in Twitter messages', in *Proceedings of the 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference. Stanford University, CA, USA, May 27-31, 2014*, pp. 27–31, (2014).

[17] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht-Ecklundt, Harald C. Traue, and Henrik Kessler, 'Mapping discrete emotions into the dimensional space: An empirical approach', in *SMC 2012 — Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics. Seoul, Korea, 14-17 October 2012*, pp. 3316–3320, (2012).

[18] Phil Katz, Matthew Singleton, and Richard Wicentowski, 'SWAT-MP: The SemEval-2007 systems for Task 5 and Task 14', in SemEval-2007 *— Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 308–313, (2007).

[19] Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, and Andrés Montoyo, 'UA-ZBSA: A headline emotion classification through Web information', in SemEval-2007 *— Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 334–337, (2007).

[20] Nicolas Leveau, Sandra Jhean-Larose, Guy Denhière, and Ba-Linh Nguyen, 'Validating an interlingual metanorm for emotional analysis of texts', *Behavior Research Methods*, **44**(4), 1007–1014, (2012).

[21] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

[22] Myriam D. Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen, 'Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text', *IEEE Transactions on Affective Computing*, **5**(2), 101–111, (2014).

[23] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, 'Affect Analysis Model: Novel rule-based approach to affect sensing from text', *Natural Language Engineering*, **17**(1), 95–135, (2011).

[24] Andrew Ortony, Gerald L. Clore, and Allan M. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

[25] Bo Pang and Lillian Lee, 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135, (2008).

[26] R. W. Picard, *Affective Computing*, MIT Press, Cambridge/MA, 1997.

[27] Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman, 'Modelling valence and arousal in Facebook posts', in *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis @ NAACL-HLT 2016. San Diego, CA, USA, June 16, 2016*, pp. 9–15, (2016).

[28] James A. Russell and Albert Mehrabian, 'Evidence for a three-factor theory of emotions', *Journal of Research in Personality*, **11**(3), 273–294, (1977).

[29] Klaus R. Scherer, 'Psychological models of emotion', in *The Neuropsychology of Emotion*, ed., Joan C. Borod, 137–162, Oxford University Press, Oxford, U.K.; New York, N.Y., (2000).

[30] Hashim Sharif, Fareed Zaffar, Ahmed Abbasi, and David Zimbra, 'Detecting adverse drug reactions using a sentiment classification framework', in *Proceedings of the 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, CA, USA, May 27-31, 2014*, (2014).

[31] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng, 'Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks', in *EMNLP 2008 — Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, USA, October 25-27, 2008*, pp. 254–263, (2008).

[32] Jacopo Staiano and Marco Guerini, 'Depeche Mood: A lexicon for emotion analysis from crowd annotated news', in *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, June 23-25 2014*, volume 2: Short Papers, pp. 427–433, (2014).

[33] Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James, 'Characterization of the affective norms for English words by discrete emotional categories', *Behavior Research Methods*, **39**(4), 1020–1024, (2007).

[34] Carlo Strapparava and Rada Mihalcea, 'SemEval-2007 Task 14: Affective text', in SemEval-2007 *— Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 70–74, (2007).

[35] Carlo Strapparava and Rada Mihalcea, 'Learning to identify emotions in text', in *SAC 2008 — Proceedings of the 2008 ACM Symposium on Applied Computing. Fortaleza, Ceará, Brazil, March 16-20, 2008*, pp. 1556–1560, (2008).

[36] Carlo Strapparava and Alessandro Valitutti, 'WordNet-Affect: An affective extension of WordNet', in *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation. In Memory of Antonio Zampolli. Lisbon, Portugal, 24-30 May, 2004*, volume 4, pp. 1083–1086, (2004).

[37] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock, 'The affective weight of lexicon', in *LREC 2006 — Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006*, pp. 423–426, (2006).

[38] Tony Veale and Yanfen Hao, 'Detecting ironic intent in creative comparisons', in *ECAI 2010 — Proceedings of the 19th European Conference on Artificial Intelligence. Lisbon, Portugal, 16-20 August 2010*, eds., Helder Coelho, Rudi Studer, and Michael J. Wooldridge, number 215 in Frontiers in Artificial Intelligence and Applications, pp. 765–770, Amsterdam, The Netherlands, (2010). IOS Press.

[39] Amy Beth Warriner, Victor Kuperman, and Marc Brysbært, 'Norms of valence, arousal, and dominance for 13,915 English lemmas', *Behavior Research Methods*, **45**(4), 1191–1207, (2013).

8 EmoBank

Reference

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Author Contributions

Udo Hahn performed supervision and project administration. I performed data acquisition, methodology, model, and software development, experimental design and execution of experiments, as well as data analysis and visualization. Conception and writing were performed jointly by both authors.

EMOBANK: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis

Sven Buechel and **Udo Hahn**
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
{sven.buechel,udo.hahn}@uni-jena.de
http://www.julielab.de

Abstract

We describe EMOBANK, a corpus of 10k English sentences balancing multiple genres, which we annotated with dimensional emotion metadata in the Valence-Arousal-Dominance (VAD) representation format. EMOBANK excels with a bi-perspectival and bi-representational design. On the one hand, we distinguish between writer's and reader's emotions, on the other hand, a subset of the corpus complements dimensional VAD annotations with categorical ones based on Basic Emotions. We find evidence for the supremacy of the reader's perspective in terms of IAA and rating intensity, and achieve close-to-human performance when mapping between dimensional and categorical formats.

1 Introduction

In the past years, the analysis of affective language has become one of the most productive and vivid areas in computational linguistics. In the early days, the prediction of the semantic polarity (positiveness or negativeness) was in the center of interest, but in the meantime, research activities shifted towards a more fine-grained modeling of sentiment. This includes the extension from only two to multiple polarity classes or even real-valued scores (Strapparava and Mihalcea, 2007), the aggregation of multiple aspects of an opinion item into a composite opinion statement for the whole item (Schouten and Frasincar, 2016), and sentiment compositionality (Socher et al., 2013).

Yet, two important features of fine-grained modeling still lack appropriate resources, namely shifting towards psychologically more adequate models of emotion (Strapparava, 2016) and distinguishing between writer's *vs.* reader's perspective on emotion ascription (Calvo and Mac Kim, 2013). We close both gaps with EMOBANK, the first large-scale text corpus which builds on the Valence-Arousal-Dominance model of emotion, an approach that has only recently gained increasing popularity within sentiment analysis. EMOBANK not only excels with a genre-balanced selection of sentences, but is based on a *bi-perspectival* annotation strategy (distinguishing the emotions of writers and readers), and includes a *bi-representationally* annotated subset (which has previously been annotated with Ekman's Basic Emotions) so that mappings between both representation formats can be performed. EMOBANK is freely available for academic purposes.[1]

2 Related Work

Models of emotion are commonly subdivided into *categorical* and *dimensional* ones, both in psychology and natural language processing (NLP). Dimensional models consider affective states to be best described relative to a small number of independent emotional dimensions (often two or three): *Valence* (corresponding to the concept of polarity), *Arousal* (degree of calmness or excitement), and *Dominance*[2] (perceived degree of control over a situation); the VAD model. Formally, the VAD dimensions span a three-dimensional real-valued vector space as illustrated in Figure 1. Alternatively, categorical models, such as the six *Basic Emotions* by Ekman (1992) or the *Wheel of Emotion* by Plutchik (1980), conceptualize emotions as discrete states.[3]

In contrast to categorical models which were used early on in NLP (Ovesdotter Alm et al., 2005; Strapparava and Mihalcea, 2007), dimensional

[1] https://github.com/JULIELab/EmoBank

[2] This dimension is sometimes omitted (the VA model).

[3] Both dimensional and categorical formats allow for numerical scores regarding their dimensions/categories.

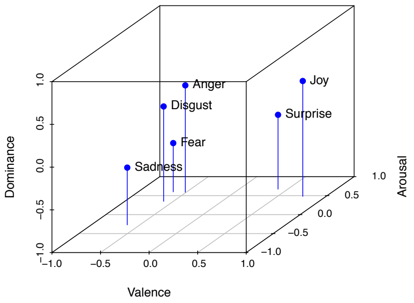


Figure 1: The affective space spanned by the three VAD dimensions. As an example, we here include the positions of Ekman's six Basic Emotions as determined by Russell and Mehrabian (1977).

models have only recently received increased attention in tasks such as word and document emotion prediction (see, e.g., Yu et al. (2015), Köper and Schulte im Walde (2016), Wang et al. (2016), Buechel and Hahn (2016)).

In spite of this shift in modeling focus, VA(D)-annotated corpora are surprisingly rare in number and small in size, and also tend to be restricted in reliability. ANET, for instance, comprises only 120 sentences designed for psychological research (Bradley and Lang, 2007), while Preoţiuc-Pietro et al. (2016) created a corpus of 2,895 English Facebook posts relying on only two annotators. Yu et al. (2016) recently presented a corpus of 2,009 Chinese sentences from various online texts.

As far as categorical models for emotion analysis are concerned, many studies use incompatible subsets of category systems, which limits their comparability (Buechel and Hahn, 2016; Calvo and Mac Kim, 2013). This also reflects the situation in psychology where there is still no consensus on a set of fundamental emotions (Sander and Scherer, 2009). Here, the VAD model has a major advantage: Since the dimensions are designed as being independent, results remain comparable dimension-wise even in the absence of others (e.g., Dominance). Furthermore, dimensional models are the predominant format for lexical affective resources in behavioral psychology as evident from the huge number of datasets available for a wide range of languages (see, e.g., Warriner et al. (2013), Stadthagen-Gonzalez et al. (2016), Moors et al. (2013) and Schmidtke et al. (2014)).

For the acquisition of VAD values from participant's self-perception, the Self-Assessment Manikin (SAM; Lang (1980), Bradley and Lang (1994)) has turned out as the most important and (to our knowledge) only standardized instrument (Sander and Scherer, 2009). SAM iconically displays differences in Valence, Arousal and Dominance by a set of anthropomorphic cartoons on a multi-point scale (see Figure 2).

While it is common for more basic sentiment analysis systems in NLP to map the many different possible interpretations of a sentence's affective meaning into a single assessment ("its sentiment"), there is an increasing interest in a more fine-grained approach where emotion expressed by writers is modeled separately from emotion evoked in readers. An utterance like "Italy defeats France in the World Cup Final" may be completely neutral from the *writer's* viewpoint (presumably a professional journalist), but is likely to evoke rather adverse emotions in Italian and French *readers* (Katz et al., 2007).

In this line of work, Tang and Chen (2012) examine the relation between the sentiment of microblog posts and the sentiment of their comments (as a proxy for reader emotion). Liu et al. (2013) model the emotion of a news reader jointly with the emotion of a comment writer using a co-training approach. This contribution was followed up by Li et al. (2016) who propose a two-view label propagation approach instead. However, to our knowledge, only Mohammad and Turney (2013) investigated the effects of these perspectives on annotation quality, finding differences in inter-annotator agreement (IAA) relative to the exact phrasing of the annotation task.

In a similar vein to the writer-reader distinction, identifying the *holder* or *source* of an opinion or sentiment also aims at describing the affective information entailed in a sentence in more detail (Wiebe et al., 2005; Seki et al., 2009). Thus, opinion statements that can directly be attributed to the writer can be distinguished from references to other's opinions. A related task, the detection of *stance*, focuses on inferring the writer's (dis)approval towards a given issue from a piece of text (Sobhani et al., 2016).

3 Corpus Design and Creation

The following criteria guided the data selection process of the EMOBANK corpus: First, complementing existing resources which focus on social media and/or review-style language (Yu et al., 2016; Quan and Ren, 2009), we decided to address several genres and domains of general English.

Corpus	Domain	Raw	Filtered
Se07	news headlines	1,250	1,192
Masc	blogs	1,378	1,336
	essays	1,196	1,135
	fiction	2,893	2,753
	letters	1,479	1,413
	newspapers	1,381	1,314
	travel guides	971	919
Sum		**10,548**	**10,062**

Table 1: Genre distribution of the raw and filtered EmoBank corpus.

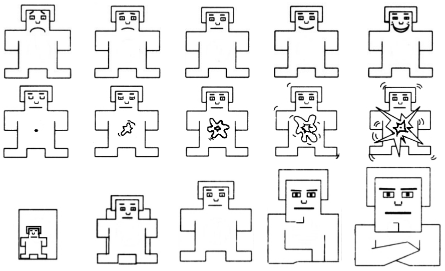


Figure 2: The modified 5-point Self-Assessment Manikin (SAM) scales for Valence, Arousal and Dominance (row-wise). Copyright of the original SAM by Peter J. Lang 1994.

Second, we conducted a pilot study on two samples (one consisting of movie reviews, the other pulled from a genre-balanced corpus) to compare the IAA resulting from different annotation perspectives (e.g., the writer's and the reader's perspective) in different domains (see Buechel and Hahn (2017) for details). Since we found differences in IAA but the results remained inconclusive, we decided to annotate the whole corpus *bi-perspectivally*, i.e., each sentence was rated according to both the (perceived) writer *and* reader emotion (henceforth, Writer and Reader).

Third, since many problems of comparing emotion analysis studies result from the diversity of emotion representation schemes (see Section 2), the ability to accurately map between such alternatives would greatly improve comparability across systems and boost the reusability of resources. Therefore, at least parts of our corpus should be annotated *bi-representationally* as well, complementing dimensional VAD ratings with annotations according to a categorical emotion model.

Following these criteria, we composed our corpus out of several categories of the *Manually Annotated Sub-Corpus of the American National Corpus* (Masc; Ide et al. (2008), Ide et al. (2010)) and the corpus of SemEval-2007 Task 14 *Affective Text* (Se07; Strapparava and Mihalcea (2007)). Masc is already annotated on various linguistic levels. Hence, our work will allow for research at the intersection of emotion and other language phenomena. Se07, on the other hand, bears annotations according to Ekman's six Basic Emotion (see Section 2) on a $[0, 100]$ scale, respectively. This collection of raw data comprises 10,548 sentences (see Table 1).

Given this large volume of data, we opted for a crowdsourcing approach to annotation. We chose CrowdFlower (Cf) over Amazon Mechanical Turk (Amt) for its quality control mechanisms and accessibility (customers of Amt, but not Cf, must be US-based). Cf's main quality control mechanism rests on *gold questions*, items for which the acceptable ratings have been previously determined by the customer. These questions are inserted into a task to restrict the workers to those performing trustworthily. We chose these gold items by automatically extracting highly emotional sentences from our raw data according to JEmAS[4], a lexicon-based tool for VAD prediction (Buechel and Hahn, 2016). The acceptable ratings were determined based on manual annotations by three students trained in linguistics. The process was individually performed for Writer and Reader with different annotators.

For each of the two perspectives, we launched an independent task on Cf. The instructions were based on those by Bradley and Lang (1999) to whom most of the VAD resources developed in psychology refer (see Section 2). We changed the 9-point SAM scales to 5-point scales (see Figure 2) in order to reduce the cognitive load during decision making for crowdworkers. For the writer's perspective, we presented a number of linguistic clues supporting the annotators in their rating decisions, while, for the reader's perspective, we asked what emotion would be evoked in an *average* reader (rather than asking for the rater's personal feelings). Both adjustments were made to establish more objective criteria for the exclusion of untrustworthy workers. We provide the instructions along with our dataset.

For each sentence, five annotators generated VAD ratings. Thus, a total of 30 ratings were gathered per sentence (five ratings for each of the three VAD dimensions and two annotation perspectives, Writer and Reader). Ten sentences were presented at a time. The task was available for work-

[4] `https://github.com/JULIELab/JEmAS`

ers located in the UK, the US, Ireland, Canada, Australia or New Zealand. The total annotation costs amounted to $1,578.

Upon inspection of the individual judgments, we found that the VAD rating $(1, 1, 1)$ was heavily overrepresented. We interpret this skewed coding distribution as a bias mainly due fraudulent responses since, from a psychological view, this rating is highly improbable (Warriner et al., 2013). Accordingly, we decided to remove all of these ratings (about 10% for each of the tasks; the 'Filtered' condition in Table 1) because these annotations would have inserted a systematic bias into our data which we consider more harmful than erroneously removing a few honest outliers. For each sentence with two or more remaining judgments, its final emotion annotation is determined by averaging these valid ratings leading to a total of 10,062 sentences bearing VAD values for *both* perspectives (see Table 1).

This makes EmoBank to the best of our knowledge by far the largest corpus for dimensional emotion models and, with the exception of the dataset by Quan and Ren (2009) (which is problematic in having only *one* annotator per sentence), the largest gold standard for any emotion format (both dimensional and categorical). Even compared with polarity corpora it is still reasonably large (e.g., similar in size to the *Stanford Sentiment Treebank* (Socher et al., 2013)).

4 Analysis and Results

For continuous, real-valued numbers, well-known metrics for IAA, such as Cohen's κ or F-score, are inappropriate as these are designed for nominally scaled variables. Instead, Pearson's correlation coefficient (r) or Mean Absolute Error (*MAE*) are often applied for this setting (Strapparava and Mihalcea, 2007; Yu et al., 2016). Accordingly, for each annotator, we compute r and *MAE* between their own and the aggregated EmoBank annotation and average these values for each VAD dimension. This results in one IAA value per metric (r or *MAE*), perspective and dimension (Table 2).

As average over the VAD dimensions, we achieve a satisfying IAA of $r > .6$ for both perspectives. The Reader results in significantly higher correlation,[5] but also higher error than Writer ($p < .05$ for Valence in r and for all dimensions in *MAE* using a two-tailed t-test).

	Valence	Arousal	Dominance	**Av.**
r_{writer}	0.698	0.578	0.540	0.605
r_{reader}	0.738	0.595	0.570	0.634
MAE_{writer}	0.300	0.388	0.316	0.335
MAE_{reader}	0.349	0.441	0.367	0.386

Table 2: IAA for the three VAD dimensions.

Prior work found that a large portion of language may actually be neutral in terms of emotion (Ovesdotter Alm et al., 2005). However, a too narrow rating distribution (i.e., most of the ratings being rather neutral relative to the three VAD dimensions) may be a disadvantageous property for training data. Therefore, we regard the *emotionality* of ratings as another quality criterion for emotion annotation complementary to IAA.

We capture this notion as the absolute difference of a sentence's aggregated rating from the neutral rating (3, in our case), averaged over all VAD dimensions. Comparing the average emotionality of all sentences between Writer and Reader, we find that the latter perspective also excels with significantly higher emotionality than the Writer ($p < .001$; two-tailed t-test).

These beneficial characteristics of the Reader perspective (better correlation-based IAA and emotionality) contrast with its worse error-based IAA. Thus, we decided to examine the relationship between error and emotionality between the two perspectives more closely: Let V, A, D be three $m \times n$-matrices where m corresponds to the number of sentences and n to the number of annotators so that the three matrices yield all the individual ratings for Valence, Arousal and Dominance, respectively. Then we define the *sentence-wise error* for sentence i (SWE$_i$) as

$$\text{SWE}_i := \frac{1}{3} \sum_{X \in \{V,A,D\}} \frac{1}{n} \sum_{j=1}^{n} |\overline{X_i} - X_{ij}| \qquad (1)$$

where $\overline{X_i} := \frac{1}{n} \sum_{j=1}^{n} X_{ij}$. We compute SWE values for reader and writer perspective individually. We can now examine the dependency between error and emotionality by subtracting, for each sentence, SWE and emotionality for both perspectives from another (resulting in one *difference in error* and one *difference in emotionality* value).

Our data reveal a strong correlation ($r = .718$) between these data series, so that the more the ratings for a sentence differ in emotionality (compar-

[5] Note that using this set-up, obtaining statistical significance is very rare, since the number of cases is based on the number of raters.

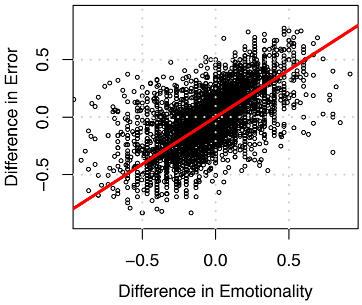


Figure 3: Differences in emotionality and differences in error between WRITER and READER, each sentence corresponding to one data point; regression line depicted in red.

ing between the perspectives), the more they differ in error as well. Running linear regression on these two data rows, we find that the regression line runs straight through the origin (intercept is *not* significantly different from 0; $p = .992$; see Figure 3). This means that without difference in emotionality, WRITER and READER rating for a sentence do, on average, *not* differ in error. Hence, our data strongly suggest that READER is the superior perspective yielding better inter-annotator *correlation* and emotionality without overproportionally increasing inter-annotator *error*.

5 Mapping between Emotion Formats

Making use of the bi-representational subset of our corpus (SE07), we now examine the feasibility of automatically mapping between dimensional and categorical models. For each Basic Emotion category, we train one k Nearest Neighbor model given all VAD values of either WRITER, READER or both combined as features. Training and hyperparameter selection was performed using 10-fold cross-validation.

Comparing the correlation between our models' predictions and the actual annotations (in categorical format) with the IAA as reported by Strapparava and Mihalcea (2007), we find that this approach already comes close to human performance (see Table 3). Once again, READER turns out to be superior in terms of the achieved mapping performance compared to WRITER. However, both perspectives combined yield even better results. In this case, our models' correlation with the actual SE07 rating is as good as or even better than the average human agreement. Note that the SE07 ratings are in turn based on averaged human judgments. Also, the human IAA differs a lot between the Basic Emotions and is even $r < .5$ for Disgust and Surprise. For the four categories with a reasonable IAA, Joy, Anger, Sadness and Fear, our best models, on average, actually outperform human agreement. Thus, our data shows that automatically mapping between representation formats is feasible at a performance level on par with or even surpassing human annotation capability. This finding suggests that, for a dataset with high-quality annotations for one emotion format, automatic mappings to another format may be just as good as creating these new annotations by manual rating.

	Joy	Ang	Sad	Fea	Dsg	Srp	**Av.**
IAA	.60	.50	.68	.64	.45	.36	.54
W	.68	.40	.67	.47	.27	.15	.44
R	.73	.47	.68	.54	.36	.15	.49
WR	.78	.50	.74	.56	.36	.17	.52
D_W	+.08	−.10	−.01	−.17	−.17	−.21	−.09
D_R	+.13	−.03	+.00	−.10	−.09	−.22	−.05
D_{WR}	+.18	+.00	+.05	−.08	−.09	−.19	−.02

Table 3: IAA by Strapparava and Mihalcea (2007) compared to mapping performance of KNN models using writer's, reader's or both's VAD scores as features (W, R and WR, respectively), both in Pearson's r. Bottom section: difference of respective model performance (W, R and WR) and IAA.

6 Conclusion

We described the creation of EMOBANK, the first large-scale corpus employing the dimensional VAD model of emotion and one of the largest gold standards for *any* emotion format. This genre-balanced corpus is also unique for having two kinds of double annotations. First, we annotated for both writer and reader emotion; second, for a subset of the EMOBANK, ratings for categorical Basic Emotions as well as VAD dimensions are now available. The statistical analysis of our corpus revealed that the reader perspective yields both better IAA values and more emotional ratings. For the bi-representationally annotated subcorpus, we showed that an automatic mapping between categorical and dimensional formats is feasible with near-human performance using standard machine leraning techniques.

Acknowledgments

We thank The Center for the Study of Emotion and Attention, University of Florida, for granting us access to the Self-Assessment-Manikin (SAM).

References

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Margaret M. Bradley and Peter J. Lang. 2007. Affective norms for English text (ANET): Affective ratings of text and instruction manual. Technical Report D-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In Gal A. Kaminka, Maria Fox, Paolo Bouquet, Eyke Hüllermeier, Virginia Dignum, Frank Dignum, and Frank van Harmelen, editors, *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS 2016). The Hague, The Netherlands, August 29 - September 2, 2016*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 1114–1122, Amsterdam, Berlin, Washington, D.C. IOS Press.

Sven Buechel and Udo Hahn. 2017. Readers *vs.* writers *vs.* texts: Coping with different perspectives of text understanding in emotion annotation. In *LAW 2017 — Proceedings of the 11th Linguistic Annotation Workshop. Valencia, Spain, April 3, 2017*.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

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

Nancy C. Ide, Collin F. Baker, Christiane Fellbaum, Charles J. Fillmore, and Rebecca J. Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan E. J. M. Odijk, Stelios Piperidis, and Daniel Tapias, editors, *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, 26 May - June 1, 2008*, pages 2455–2461.

Nancy C. Ide, Collin F. Baker, Christiane Fellbaum, and Rebecca J. Passonneau. 2010. The Manually Annotated Sub-Corpus: A community resource for and by the people. In Jan Hajič, M. Sandra Carberry, and Stephen Clark, editors, *ACL 2010 — Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 11-16 July 2010*, volume 2: Short Papers, pages 68–73.

Phil Katz, Matthew Singleton, and Richard Wicentowski. 2007. SWAT-MP: The SemEval-2007 systems for Task 5 and Task 14. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *SemEval-2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007. Prague, Czech Republic, June 23-24, 2007*, pages 308–313.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan E. J. M. Odijk, and Stelios Piperidis, editors, *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 2595–2598.

Peter J. Lang. 1980. Behavioral treatment and biobehavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, and T. A. Williams, editors, *Technology in Mental Health Care Delivery Systems*, pages 119–137. Ablex, Norwood/NJ.

Shoushan Li, Jian Xu, Dong Zhang, and Guodong Zhou. 2016. Two-view label propagation to semi-supervised reader emotion classification. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan, December 11-16, 2016*, volume Technical Papers, pages 2647–2655.

Huanhuan Liu, Shoushan Li, Guodong Zhou, Chu-Ren Huang, and Peifeng Li. 2013. Joint modeling of news reader's and comment writer's emotions. In Hinrich Schütze, Pascale Fung, and Massimo Poesio, editors, *ACL 2013 — Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, August 4-9, 2013*, volume 2: Short Papers, pages 511–515.

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

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45(1):169–177.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In Raymond J. Mooney, Christopher Brew, Lee-Feng Chien, and Katrin Kirchhoff, editors, *HLT-EMNLP 2005 — Proceedings of the Human Language Technology Conference & 2005 Conference on Empirical*

Methods in Natural Language Processing. Vancouver, British Columbia, Canada, 6-8 October 2005, pages 579–586.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research and Experience*, 1(3):3–33.

Daniel Preoţiuc-Pietro, Hansen Andrew Schwartz, Gregory Park, Johannes C. Eichstaedt, Margaret L. Kern, Lyle H. Ungar, and Elizabeth P. Shulman. 2016. Modelling valence and arousal in Facebook posts. In Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andrés Montoyo, editors, *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, pages 9–15.

Changqin Quan and Fuji Ren. 2009. Construction of a blog emotion corpus for Chinese emotional expression analysis. In Philipp Koehn and Rada Mihalcea, editors, *EMNLP 2009 — Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. A Meeting of SIGDAT, a Special Interest Group of ACL @ ACL-IJCNLP 2009. Singapore, 6-7 August 2009*, pages 1446–1454.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

David Sander and Klaus R. Scherer, editors. 2009. *The Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press, Oxford; New York.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. 2014. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, 46(4):1108–1118.

Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Yohei Seki, Noriko Kando, and Masaki Aono. 2009. Multilingual opinion holder identification using author and authority viewpoints. *Information Processing & Management*, 45(2):189–199.

Parinaz Sobhani, Saif M. Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In Claire Gardent, Raffaella Bernardi, and Ivan Titov, editors, **SEM 2016 — Proceedings of the 5th Joint Conference on Lexical and Computational Semantics @ ACL 2016. Berlin, Germany, August 11-12, 2016*, pages 159–169.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Timothy Baldwin and Anna Korhonen, editors, *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA, 18-21 October 2013*, pages 1631–1642.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez Sánchez, and Marc Brysbaert. 2016. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*. 10.3758/s13428-015-0700-2.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *SemEval-2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007. Prague, Czech Republic, June 23-24, 2007*, pages 70–74.

Carlo Strapparava. 2016. Emotions and NLP: Future directions. In Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andrés Montoyo, editors, *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, page 180.

Yi-jie Tang and Hsin-Hsi Chen. 2012. Mining sentiment words from microblogs for predicting writer-reader emotion transition. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan E. J. M. Odijk, and Stelios Piperidis, editors, *LREC 2012 — Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012*, pages 1226–1229.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In Antal van den Bosch, Katrin Erk, and Noah A. Smith, editors, *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 2: Short Papers, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Janyce M. Wiebe, Theresa Ann Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3 (Special Issue on “Advances in Question Answering”)):165–210.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In Yuji Matsumoto, Chengqing Zong, and Michael Strube, editors, *ACL-IJCNLP 2015 — Proceedings of the 53rd*

Annual Meeting of the Association for Computational Linguistics & 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, July 26-31, 2015, volume 2: Short Papers, pages 788–793.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In Kevin C. Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 540–545.

9 Modeling Empathy and Distress in Reaction to News Stories

Reference

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

Author Contributions

Anneke Buffone performed supervision, project administration, methodology development and experimental design. Barry Slaff performed data analysis and visualization. Lyle Ungar performed supervision and project administration. João Sedoc performed supervision and data acquisition. I performed data acquisition, model and software development, experimental design and execution of experiments, data analysis and visualization. Conception and writing were jointly performed by all authors. See also contribution statement in footnote "*" on the following page.

Modeling Empathy and Distress in Reaction to News Stories

Sven Buechel *†3 **Anneke Buffone** *1 **Barry Slaff** 1 **Lyle Ungar** 1,2 **João Sedoc** 1,2

1 Positive Psychology Center, University of Pennsylvania
2 Computer & Information Science, University of Pennsylvania
3 JULIE Lab, Friedrich-Schiller-Universität Jena

https://wwbp.org 1,2 https://julielab.de 3

Abstract

Computational detection and understanding of empathy is an important factor in advancing human-computer interaction. Yet to date, text-based empathy prediction has the following major limitations: It underestimates the psychological complexity of the phenomenon, adheres to a weak notion of ground truth where empathic states are ascribed by third parties, and lacks a shared corpus. In contrast, this contribution presents the first publicly available gold standard for empathy prediction. It is constructed using a novel annotation methodology which reliably captures empathy assessments by the writer of a statement using multi-item scales. This is also the first computational work distinguishing between multiple forms of empathy, empathic concern, and personal distress, as recognized throughout psychology. Finally, we present experimental results for three different predictive models, of which a CNN performs the best.

1 Introduction

Over two decades after the seminal work by Picard (1997) the quest of *Affective Computing*, to ease the interaction with computers by giving them a sense of how emotions shape our perception and behavior, is still far from being fulfilled. Undoubtedly, major progress has been made in NLP, with sentiment analysis being one of the most vivid and productive areas in recent years (Liu, 2015).

However, the vast majority of contributions has focused on *polarity prediction*, typically only distinguishing between positive and negative feeling or evaluation, usually in social media postings or product reviews (Rosenthal et al., 2017; Socher et al., 2013). Only very recently, researchers started exploring more sophisticated models of human emotion on a larger scale (Wang et al., 2016; Abdul-Mageed and Ungar, 2017; Mohammad and Bravo-Marquez, 2017a; Buechel and Hahn, 2017, 2018a,b). Yet such approaches, often rooted in psychological theory, also turned out to be more challenging in respect to annotation and modeling (Strapparava and Mihalcea, 2007).

Surprisingly, one of the most valuable affective phenomena for improving human-machine interaction has received surprisingly little attention: *Empathy*. Prior work focused mostly on *spoken dialogue*, commonly addressing conversational agents, psychological interventions, or call center applications (McQuiggan and Lester, 2007; Fung et al., 2016; Pérez-Rosas et al., 2017; Alam et al., 2017).

In contrast, to the best of our knowledge, only three contributions (Xiao et al., 2012; Gibson et al., 2015; Khanpour et al., 2017) previously addressed *text-based* empathy prediction[1] (see Section 4 for details). Yet, all of them are limited in three ways: (a) neither of their corpora are available leaving the NLP community without shared data, (b) empathy ratings were provided by others than the one actually experiencing it which qualifies only as a weak form of ground truth, and (c) their notion of empathy is quite basic, falling short of current and past theory.

* These authors contributed equally to this work. Anneke Buffone designed and supervised the crowdsourcing task and the survey described in Section 2, and provided psychological background knowledge. Sven Buechel was responsible for corpus creation, data analysis, and modeling. The technical set-up of the crowdsourcing task and the survey was done jointly by both first authors.

†Work conducted while being at the University of Pennsylvania.

[1] Psychological studies commonly distinguish between *state* and *trait* empathy. While the former construct describes the amount of empathy a person experiences as a direct result of encountering a given stimulus, the latter refers to how empathetic one is on average and across situations. This studies exclusively addresses *state empathy*. For a contribution addressing *trait empathy* from an NLP perspective, see Abdul-Mageed et al. (2017).

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4758–4765
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

In this contribution we present the first publicly available gold standard for text-based empathy prediction. It is constructed using a novel annotation methodology which reliably captures empathy assessments via multi-item scales. The corpus as well as our work as a whole is also unique in being—to the best of our knowledge—the first computational approach differentiating *multiple types of empathy*, empathic concern and personal distress, a distinction well recognized throughout psychology and other disciplines.[2]

2 Corpus Design and Methodology

Background. Most psychological theories of empathic states are focused on reactions to negative rather than positive events. Empathy for positive events remains less well understood and is thought to be regulated differently (Morelli et al., 2015). Thus we focus on empathetic reactions to need or suffering. Despite the fact that everyone has an immediate, implicit understanding of empathy, research has been vastly inconsistent in its definition and operationalization (Cuff et al., 2016). There is agreement, however, that there are multiple forms of empathy (see below). The by far most widely cited state empathy scale is Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987), henceforth *empathy* and *distress*.

Distress is a self-focused, negative affective state that occurs when one feels upset due to witnessing an entity's suffering or need, potentially via "catching" the suffering target's negative emotions. Empathy is a warm, tender, and compassionate feeling for a suffering target. It is other-focused, retains self-other separation, and is marked by relatively more positive affect (Batson and Shaw, 1991; Goetz et al., 2010; Mikulincer and Shaver, 2010; Sober and Wilson, 1997).

Selection of News Stories. Two research interns (psychology undergraduates) collected a total of 418 articles from popular online news platforms, selected to likely evoke empathic reactions, after being briefed on the goal and background of this study. These articles were then used to elicit empathic responses in participants.

Acquiring Text and Ratings. The corpus acquisition was set up as a crowdsourcing task on `MTurk.com` pointing to a `Qualtrics.com` questionnaire. The participants completed background measures on demographics and personality, and then proceeded to the main part of the survey where they read a random selection of five of the news articles. After reading each of the articles, participants were asked to rate their level of empathy and distress before describing their thoughts and feelings about it in writing.

In contrast to previous work, this set-up allowed us to acquire empathy scores of the actual *writer* of a text, instead of having to rely on an external evaluation by third parties (often student assistants with background in computer science). Arguably, our proposed annotation methodology yields more appropriate gold data, yet also leads to more variance in the relationship between linguistic features and empathic state ratings. That is because each rating reflects a single individual's feelings rather than a more stable average assessment by multiple raters. To account for this, we use *multi-item scales* as is common practice in psychology. I.e., participants give ratings for multiple items measuring the same construct (e.g., empathy) which are then averaged to obtain more reliable results. As far as we know, this is the first time that multi-item scales are used in sentiment analysis.[3]

In our case, participants used Batson's Empathic Concern – Personal Distress Scale (see above), i.e, rating 6 items for empathy (e.g., *warm, tender, moved*) and 8 items for distress (e.g., *troubled, disturbed, alarmed*) using a 7-point scale for each of those (see Appendix for details). After rating their empathy, participants were asked to share their feelings about the article as they would with a friend in a private message or with a group of friends as a social media post in 300 to 800 characters. Our final gold standard consists of these *messages* combined with the numeric ratings for empathy and distress.

In sum, 403 participants completed the survey. Median completion time was 32 minutes and each participant received 4 USD as compensation.

Post-Processing. Each message was manually reviewed by the authors. Responses which deviated from the task description (e.g., mere copying from the articles at display) were removed (31 responses, 155 messages), leading to a total 1860 messages in our final corpus. Gold ratings for empathy and distress were derived by averaging the respective items of the two multi-item scales.

[2] Data and code are available at: `https://github.com/wwbp/empathic_reactions`

[3] Here, we use *sentiment* as an umbrella term subsuming semantic orientation, emotion, as well as highly related concepts such as empathy.

	E	D	Message
(1)	4.8	3.1	*I'm sorry to hear that about Dakota's parents. Even when you are adult it must be hard to see your parents splitting up. No one wants that to happen and it's unfortunate that her parents couldn't work it out. I hope they are able to still remain civil around the kids and family. Just because it didn't work romantically doesn't mean it won't work at all.*
(2)	4.0	5.5	*Here's an article about crazed person who murdered two unfortunate women overseas. Life is crazy. I can't imagine what the families are going through. Having to go to or being forced into sex work is bad enough, but for it to end like this is just sad. It feels like there's no place safe in this world to be a woman sometimes.*
(3)	1.0	1.3	*I just read an article about some chowder-head who used a hammer and a pick ax to destroy Donald Trump's star on the Hollywood walk of fame. Wow, what a great protest. You sure showed him. Good job. Lol, can you believe this garbage? Who has such a hollow and pathetic life that they don't have anything better to do with their time than commit petty vandalism because they dislike some politician? What a dingus.*

Table 1: Illustrative examples from our newly created gold standard with ratings for empathy (**E**) and distress (**D**).

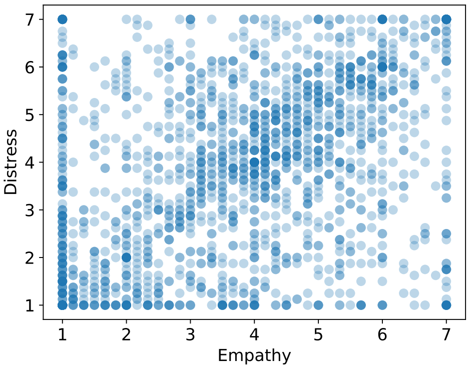


Figure 1: Scatter plot of the bivariate distribution of empathy and distress ratings.

3 Corpus Analysis

For a first impression of the language of our new gold standard, we provide illustrative examples in Table 1. The participant in Example (1) displays higher empathy than distress, (2) displays higher distress than empathy, and (3) shows neither empathic state, but employs sarcasm, colloquialisms and social-media-style acronyms to express lack of emotional response to the article. As can be seen, the language of our corpus is diverse and authentic, featuring many phenomena of natural language which render its computational understanding difficult, thus constituting a sound but challenging gold standard for empathy prediction.

Token Counts. We tokenized the 1860 messages using NLTK tools (Bird, 2006). In total, our corpus amounts to $173,686$ tokens. Individual message length varies between 52 and 198 tokens, the median being 84. See Appendix for details.

Rating Distribution. Figure 1 displays the bivariate distribution of empathy and distress ratings. As can be seen both target variables have a clear linear dependence, yet show only a moderate Pearson correlation of r=$.451$, similar to what was found in prior research (Batson et al., 1987, 1997). This finding supports that the two scales capture distinct affective phenomena and underscores the importance of our decision to describe empathic states in terms of *multiple* target variables, constituting a clear advancement over previous work. Both kinds of ratings show good coverage over the full range of the scales.

Reliability of Ratings. Since each message is annotated by only one rater, its author, typical measures of inter-rater agreement are not applicable. Instead, we compute *split-half reliability* (SHR), a standard approach in psychology (Cronbach, 1947) which also becomes increasingly popular in sentiment analysis (Mohammad and Bravo-Marquez, 2017a; Buechel and Hahn, 2018a). SHR is computed by splitting the ratings for the individual scale items (e.g., *warm*, *tender*, etc. for empathy) of all participants randomly into two groups, averaging the individual item ratings for each group and participant, and then measuring the correlation between both groups. This process is repeated 100 times with random splits, before again averaging the results. Doing so for empathy and distress, we find very high[4] SHR values of r=.875 and .924, respectively.

4 Modeling Empathy and Distress

In this section, we provide experimental results for modeling empathy and distress ratings based on the participants' messages (see Section 2). We examine three different types of models, varying in

[4] For a comparison against previously reported SHR values for different emotional categories, see Mohammad and Bravo-Marquez (2017b).

design complexity. Distinct models were trained for empathy and distress prediction.

First, ten percent of our newly created gold standard were randomly sampled to be used in development experiments. Then, the main experiment was conducted using 10-fold cross-validation (CV), providing each model with identical train-test splits to increase reliability. The dev set was excluded for the CV experiment.

Model performance is measured in terms of Pearson correlation r between predicted values and the human gold ratings. Thus, we phrase the prediction of empathy and distress as regression problems.

The input to our models is based on word embeddings, namely the publicly available FastText embeddings which were trained on Common Crawl ($\approx$600B tokens) (Bojanowski et al., 2017; Mikolov et al., 2018).

Ridge. Our first approach is Ridge regression, an ℓ^2-regularized version of linear regression. The centroid of the word embeddings of the words in a message is used as features (embedding centroid). The regularization coefficient α is automatically chosen from $\{1, .5, .1, ..., .0001\}$ during training.

FFN. Our second approach is a Feed-Forward Net with two hidden layers (256 and 128 units, respectively) with ReLU activation. Again, the embedding centroid is used as features.

CNN. The last approach is a Convolutional Neural Net.[5] We use a single convolutional layer with filter sizes 1 to 3, each with 100 output channels, followed by an average pooling layer and a dense layer of 128 units. ReLUs were used for the convolutional and again for the dense layer.

Both deep learning models were trained using the Adam optimizer (Kingma and Ba, 2015) with a fixed learning rate of 10^{-3} and a batch size of 32. We trained for a maximum of 200 epochs yet applied early stopping if the performance on the validation set did not improve for 20 consecutive epochs. We applied dropout with probabilities of .2, .5 and .5 on input, dense and pooling layers, respectively. Moreover ℓ^2 regularization of .001 was applied to the weights of conv and dense layers. Word embeddings were not updated.

The results are provided in Table 2. As can be seen, all of our models achieve satisfying performance figures ranging between r=.379 and .444, given the assumed difficulty of the task (see Section 3). On average over the two target variables, the CNN performs best, followed by Ridge and the FFN. While the CNN significantly outperforms the other models in every case, the differences between Ridge and the FFN are not statistically significant for either empathy or distress.[6] The improvements of the CNN over the other two approaches are much more pronounced for distress than for empathy. Since only the CNN is able to capture semantic effects from composition and word order, our data suggest that these phenomena are more important for predicting distress, whereas lexical features alone already perform quite well for empathy.

	Empathy	**Distress**	**Mean**
Ridge	.385	.410	.398
FFN	.379	.401	.390
CNN	**.404***	**.444***	**.424***

Table 2: Model performance for predicting empathy and distress in Pearson's r; with row-wise mean; best result per column in bold, significant ($p < .05$) improvement over other models marked with '*'.

Discussion. In comparison to closely related tasks such as emotion prediction (Mohammad and Bravo-Marquez, 2017a) our performance figures for empathy and distress prediction are generally lower. However, given the small amount of previous work for the problem at hand, we argue that our results are actually quite strong. This becomes obvious, again, in comparison with emotion analysis where early work achieved correlation values around r=.3 at most (Strapparava and Mihalcea, 2007). Yet state-of-the-art performance literally doubled over the last decade (Beck, 2017), in part due to much larger training sets.

Comparison to the limited body of previous work in text-based empathy prediction is difficult for a number of reasons, e.g., differences in domain, evaluation metric, as well as methodology and linguistic level of annotation. Khanpour et al. (2017) annotate and model empathy in online health communities on the *sentence*-level, whereas the instances in our corpus are much longer and comprise multiple sentences. In contrast to our work, they treat empathy prediction as a classification problem. Their best performing model, a CNN-LSTM, achieves an F-score of .78. Gibson

[5] Recurrent models did not perform well during development due to high sequence length.

[6] We use a two-tailed t-test for paired samples based on the results of the individual CV runs; $p < .05$.

et al. (2015) predict therapists' empathy in motivational interviews. Each therapy session transcript received one numeric score. Thus, each prediction is based on much more language data than our individual messages comprise. Their best model achieves a Spearman rank correlation of .61 using n-gram and psycholinguistic features.

Our contribution goes beyond both of these studies by, first, enriching empathy prediction with personal distress and, second, by annotating and modeling the empathic state actually felt by the writer, instead of relying on external assessments.

5 Conclusion

This contribution was the first to attempt empathy prediction in terms of *multiple* target variables, empathic concern and personal distress. We proposed a novel annotation methodology capturing empathic states actually felt by the author of a statement, instead of relying on third-party assessments. To ensure high reliability in this single-rating setting, we employ multi-item scales in line with best practices in psychology. Hereby we create the first publicly available gold standard for empathy prediction in written language, our survey being set-up and supervised by an expert psychologist. Our analysis shows that the data set excels with high rating reliability and an authentic and diverse language, rich of challenging phenomena such as sarcasm. We provide experimental results for three different predictive models, our CNN turning out superior.

Acknowledgments

Sven Buechel would like to thank his doctoral advisor Udo Hahn, JULIE Lab, for funding his research visit at the University of Pennsylvania.

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Johannes C Eichstaedt, and Lyle H Ungar. 2017. Recognizing pathogenic empathy in social media. In *ICWSM 2017 — Proceedings of the 11th International Conference on Web and Social Media*, pages 448–451, Montreal, Canada, May 15–18, 2017.

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, long papers, pages 718–728, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2017. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, pages 40–61.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

C Daniel Batson, Karen Sager, Eric Garst, Misook Kang, Kostia Rubchinsky, and Karen Dawson. 1997. Is empathy-induced helping due to self–other merging? *Journal of personality and social psychology*, 73(3):495.

C Daniel Batson and Laura L Shaw. 1991. Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological inquiry*, 2(2):107–122.

Daniel Beck. 2017. Modelling representation noise in emotion analysis using gaussian processes. In *IJCNLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 2, short papers, pages 140–145, Taipei, Taiwan, November 27 – December 1, 2017.

Steven Bird. 2006. NLTK: The natural language toolkit. In *COLING-ACL 2006 — Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, volume 4, interactive presentation sessions, pages 69–72, Sydney, Australia, July 17–21, 2006.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, short papers, pages 578–585, Valencia, Spain, April 3–7, 2017.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, volume 1, technical papers, pages 2892–2904, Santa Fe, New Mexico, USA, August 20–26, 2018.

Sven Buechel and Udo Hahn. 2018b. Word emotion induction for multiple languages as a deep multitask learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, volume 1, long papers, pages 1907–1918, New Orleans, Louisiana, USA, June 1–6, 2018.

Lee J Cronbach. 1947. Test reliability: Its meaning and determination. *Psychometrika*, 12(1):1–16.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: a review of the concept. *Emotion Review*, 8(2):144–153.

Mark H Davis. 1980. *Interpersonal Reactivity Index*. Edwin Mellen Press.

ED Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. 1985. The satisfaction with life scale. *Journal of Personality Assessment*, 49(1):71–75.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the supergirl: An empathetic personality recognition system. In *NAACL 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 3, demonstrations, pages 87–91, San Diego, California, USA, June 12–17, 2016.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Interspeech 2015 — Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pages 1947–1951, Dresden, Germany, September 6–10, 2015.

Jennifer L Goetz, Dacher Keltner, and Emiliana Simon-Thomas. 2010. Compassion: an evolutionary analysis and empirical review. *Psychological Bulletin*, 136(3):351.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *IJCNLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 2, short papers, pages 246–251, Taipei, Taiwan, November 27 – December 1, 2017.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015 — Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15, San Diego, California, USA, May 7–9, 2015.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*. Cambridge University Press.

Scott W McQuiggan and James C Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies*, 65(4):348–360.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 52–55, Miyazaki, Japan, May 7-12, 2018.

M. Mikulincer and P. R. Shaver, editors. 2010. *Prosocial motives, emotions, and behavior: The better angels of our nature*. American Psychological Association.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. WASSA-2017 shared task on emotion intensity. In *WASSA 2017 — Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP*, pages 34–49, Copenhagen, Denmark, September 8, 2017.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017b. Emotion intensities in tweets. In **SEM 2017 — Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 65–77, Vancouver, British Columbia, Canada, August 3–4, 2017.

Sylvia A Morelli, Matthew D Lieberman, and Jamil Zaki. 2015. The emerging study of positive empathy. *Social and Personality Psychology Compass*, 9(2):57–68.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, long papers, pages 1426–1435, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

R. W. Picard. 1997. *Affective Computing*. MIT Press.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *SemEval 2017 — Proceedings of the 11th International Workshop on Semantic Evaluation @ ACL*, pages 502–518, Vancouver, British Columbia, Canada, August 3–4, 2017.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. DLATK: Differential language analysis toolkit. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, volume 2, system demonstrations, pages 55–60, Copenhagen, Denmark, September 7–11, 2017.

Elliott Sober and David Sloan Wilson. 1997. Unto others: The evolution of altruism. *Harvard University.*

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 18–21, 2013.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval 2007 Task 14: Affective text. In *SemEval 2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007*, pages 70–74, Prague, Czech Republic, June 23–24, 2007.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, short papers, pages 225–230, Berlin, Germany, August 7–12, 2016.

Bo Xiao, Dogan Can, Panayiotis G Georgiou, David Atkins, and Shrikanth S Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *APSIPA 2012 — Proceedings of the 2012 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4, Hollywood, California, USA, December 3–6, 2012.

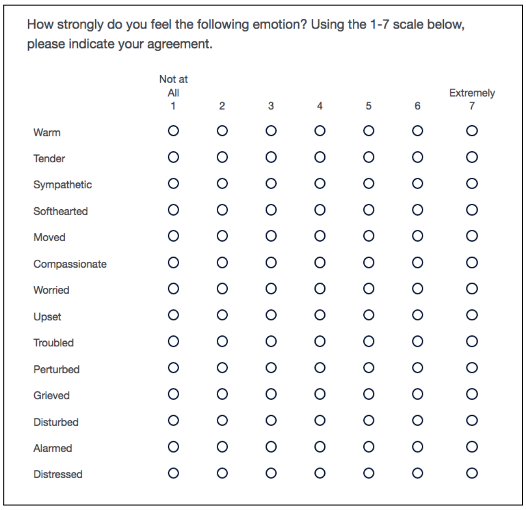
A Supplemental Material

Details on Stimulus and Instructions

Before being used in our survey, the selected news articles were categorized by the research interns who gathered them in terms of their intensity of suffering (major or minor), cause of suffering (political, human, nature or other), patient of suffering (humans, animals, environment, or other) and scale of suffering (individual or mass). Research interns also provided a short list of key words for each article. This additional information was gathered to examine the influence of these factors on empathy elicitation and modeling performance in later studies.

At the beginning of the survey participants completed background items covering general demographics (including age, gender, and ethnicity), the most commonly used *trait* empathy scale, the Interpersonal Reactivity Index (Davis, 1980), a brief assessment of the Big 5 personality traits (Gosling et al., 2003), life satisfaction (Diener et al., 1985), as well as a brief measure of generalized trust.

After reading each of the articles, participants rated their level of empathic concern and personal distress using multi-item scales. **Figure 2** shows a cropped screenshot of the survey hosted on `Qualtrics.com`. The first six items (*warm, tender, sympathetic, softhearted, moved*, and *compassionate*) refer to empathy. The last eight items (*worried, upset, troubled, perturbed, grieved, disturbed, alarmed*, and *distressed*) refer to distress.



How strongly do you feel the following emotion? Using the 1-7 scale below, please indicate your agreement.

	Not at All 1	2	3	4	5	6	Extremely 7
Warm	○	○	○	○	○	○	○
Tender	○	○	○	○	○	○	○
Sympathetic	○	○	○	○	○	○	○
Softhearted	○	○	○	○	○	○	○
Moved	○	○	○	○	○	○	○
Compassionate	○	○	○	○	○	○	○
Worried	○	○	○	○	○	○	○
Upset	○	○	○	○	○	○	○
Troubled	○	○	○	○	○	○	○
Perturbed	○	○	○	○	○	○	○
Grieved	○	○	○	○	○	○	○
Disturbed	○	○	○	○	○	○	○
Alarmed	○	○	○	○	○	○	○
Distressed	○	○	○	○	○	○	○

Figure 2: Multi-item scales for empathic concern and personal distress.

After completing the rating items, participants were instructed to describe their reactions in writing as follows: *Now that you have read this article, please write a message to a friend or friends about your feelings and thoughts regarding the article you just read. This could be a private message to a friend or something you would post on social media. Please do not identify your intended friend(s) — just write your thoughts about the article as if you were communicating with them. Please use between 300 and 800 characters.*

Further Corpus Analyses

The word clouds in **Figure 3** and **Figure 4** show 1-grams of our corpus which correlate significantly (Benjamini-Hochberg corrected $p < .05$) with high empathy and high distress ratings, respectively. In the word clouds, larger size indicates higher correlation and the color scale, gray-blue-red, indicates word frequency, dark red being most prevalent. The Differential Language Analysis Toolkit (Schwartz et al., 2017) was utilized for this analysis. As can be seen, the word clouds display high face-validity, giving further evidence for the soundness of our acquisition methodology.

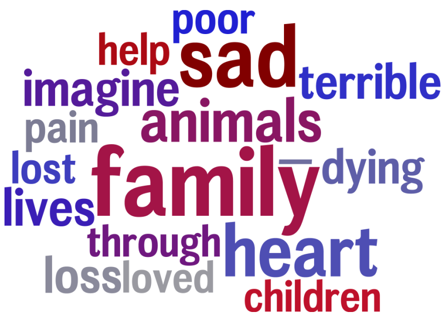


Figure 3: Word cloud of high empathy 1-grams.

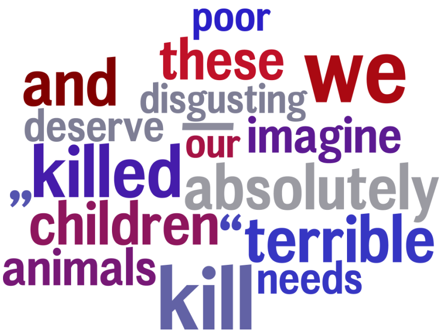


Figure 4: Word cloud of high distress 1-grams.

Figure 5 displays the distribution of the message length of our corpus in tokens. As can be seen the majority of messages contain between 60 and 100 tokens. Yet outliers go up to almost 200. The introduction of a character cap for the writing task proved successful in comparison to a pilot study where this measure has not been in place. In the latter case, the maximum number of tokens was nearly twice as high due to even stronger outliers.

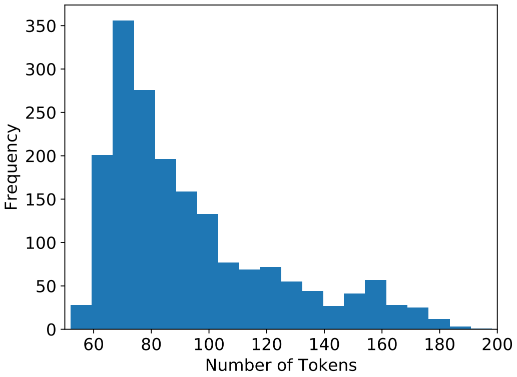


Figure 5: Histogram of message length in our corpus.

10 Word Emotion Induction as Deep Multi-Task Learning

Reference

Sven Buechel and Udo Hahn. 2018c. Word emotion induction for multiple languages as a deep multi-task learning problem. 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 1907–1918.

Author Contributions

Udo Hahn performed supervision and project administration. I performed methodology, model, and software development, experimental design and execution of experiments, as well as data analysis and visualization. Conception and writing were performed jointly by both authors.

Word Emotion Induction for Multiple Languages as a Deep Multi-Task Learning Problem

Sven Buechel & Udo Hahn
{sven.buechel|udo.hahn}@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
http://www.julielab.de

Abstract

Predicting the emotional value of lexical items is a well-known problem in sentiment analysis. While research has focused on polarity for quite a long time, meanwhile this early focus has been shifted to more expressive emotion representation models (such as Basic Emotions or Valence-Arousal-Dominance). This change resulted in a proliferation of heterogeneous formats and, in parallel, often small-sized, non-interoperable resources (lexicons and corpus annotations). In particular, the limitations in size hampered the application of deep learning methods in this area because they typically require large amounts of input data. We here present a solution to get around this language data bottleneck by rephrasing word emotion induction as a multi-task learning problem. In this approach, the prediction of each independent emotion dimension is considered as an individual task and hidden layers are *shared* between these dimensions. We investigate whether multi-task learning is more advantageous than single-task learning for emotion prediction by comparing our model against a wide range of alternative emotion and polarity induction methods featuring 9 typologically diverse languages and a total of 15 conditions. Our model turns out to outperform each one of them. Against all odds, the proposed deep learning approach yields *the largest gain* on *the smallest* data sets, merely composed of one thousand samples.

1 Introduction

Deep Learning (DL) has radically changed the rules of the game in NLP by dramatically boosting performance figures in almost all applications areas. Yet, one of the major premises of high-performance DL engines is their dependence on huge amounts of training data. As such, DL seems ill-suited for areas where training data are scarce, such as in the field of word emotion induction.

We will use the terms *polarity* and *emotion* here to distinguish between research focusing on “semantic orientation” (Hatzivassiloglou and McKeown, 1997) (the positiveness or negativeness) of affective states, on the one hand, and approaches which provide predictions based on some of the many more elaborated representational systems for affective states, on the other hand.

Originally, research activities focused on polarity alone. In the meantime, a shift towards more expressive representation models for emotion can be observed that heavily draws inspirations from psychological theory, e.g., Basic Emotions (Ekman, 1992) or the Valence-Arousal-Dominance model (Bradley and Lang, 1994).

Though this change turned out to be really beneficial for sentiment analysis in NLP, a large variety of mutually incompatible encodings schemes for emotion and, consequently, annotation formats for emotion metadata in corpora have emerged that hinder the interoperability of these resources and their subsequent reuse, e.g., on the basis of alignments or mergers (Buechel and Hahn, 2017).

As an alternative way of dealing with thus unwarranted heterogeneity, we here examine the potential of multi-task learning (MTL; Caruana (1997)) for word-level emotion prediction. In MTL for neural networks, a single model is fitted to solve multiple, independent tasks (in our case, to predict different emotional dimensions) which typically results in learning more robust and meaningful intermediate representations. MTL has been shown to greatly decrease the risk of overfitting (Baxter, 1997), work well for various NLP tasks (Setiawan et al., 2015; Liu et al., 2015; Søgaard and Goldberg, 2016; Cummins et al., 2016; Liu et al., 2017; Peng et al., 2017), and practically increases sample size, thus making it a natural choice for small-sized data sets typically found in the area of word emotion induction.

Proceedings of NAACL-HLT 2018, pages 1907–1918
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

After a discussion of related work in Section 2, we will introduce several reference methods and describe our proposed deep MTL model in Section 3. In our experiments (Section 4), we will first validate our claim that MTL is superior to single-task learning for word emotion induction. After that, we will provide a large-scale evaluation of our model featuring 9 typologically diverse languages and multiple publicly available embedding models for a total of 15 conditions. Our MTL model surpasses the current state-of-the-art for each of them, and even performs competitive relative to human reliability. Most notably however, our approach yields the largest benefit on the smallest data sets, comprising merely one thousand samples. This finding, counterintuitive as it may be, strongly suggests that MTL is particularly beneficial for solving the word emotion induction problem. Our code base as well as the resulting experimental data is freely available.[1]

2 Related Work

This section introduces the emotion representation format underlying our study and describes external resources we will use for evaluation before we discuss previous methodological work.

Emotion Representation and Data Sets. Psychological models of emotion can typically be subdivided into *discrete* (or *categorical*) and *dimensional* ones (Stevenson et al., 2007; Calvo and Mac Kim, 2013). Discrete models are centered around particular sets of emotional categories considered to be fundamental. Ekman (1992), for instance, identifies six *Basic Emotions* (Joy, Anger, Sadness, Fear, Disgust and Surprise).

In contrast, dimensional models consider emotions to be composed of several influencing factors (mainly two or three). These are often referred to as *Valence* (a positive–negative scale), *Arousal* (a calm–excited scale), and *Dominance* (perceived degree of control over a (social) situation)—the VAD model (Bradley and Lang (1994); see Figure 1 for an illustration). Many contributions though omit Dominance (the VA model) (Russell, 1980). For convenience, we will still use the term “VAD” to jointly refer to both variants (with and without Dominance).

VAD is the most common framework to acquire empirical emotion values for words in psychology.

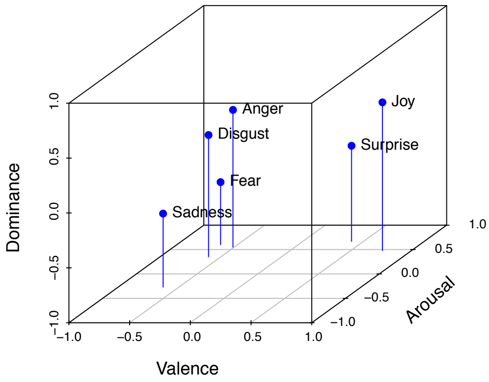


Figure 1: Affective space spanned by the Valence-Arousal-Dominance (VAD) model, together with the position of six Basic Emotions; as determined by Russell and Mehrabian (1977).

Over the years, a considerable number of such resources (also called “emotion lexicons”) have emerged from psychological research labs (as well as some NLP labs) for diverse languages. The emotion lexicons we use in our experiments are listed in Table 1. An even more extensive list of such data sets is presented by Buechel and Hahn (2018). For illustration, we also provide three sample entries from one of those lexicons in Table 2. As can be seen, the three affective dimensions behave complementary to each other, e.g., “terrorism” and “orgasm” display similar Arousal but opposing Valence.

The task we address in this paper is to predict the values for Valence, Arousal and Dominance, given a lexical item. As is obvious from these examples, we consider emotion prediction as a regression, not as a classification problem (see arguments discussed in Buechel and Hahn (2016)).

In this paper, we focus on the VAD format for the following reasons: First, note that the Valence dimension exactly corresponds to polarity (Turney and Littman, 2003). Hence, with the VAD model, emotion prediction can be seen as a generalization over classical polarity prediction. Second, to the best of our knowledge, the amount and diversity of available emotion lexicons with VAD encodings is larger than for any other format (see Table 1).

Word Embeddings. Word embeddings are dense, low-dimensional vector representations of words trained on large volumes of raw text in an unsupervised manner. The following are among today’s most popular embedding algorithms:

[1] `https://github.com/JULIELab/wordEmotions`

Source	ID	Language	Format	# Entries
Bradley and Lang (1999)	EN	English	VAD	1,034
Warriner et al. (2013)	EN+	English	VAD	13,915
Redondo et al. (2007)	ES	Spanish	VAD	1,034
Stadthagen-Gonzalez et al. (2017)	ES+	Spanish	VA	14,031
Schmidtke et al. (2014)	DE	German	VAD	1,003
Yu et al. (2016a)	ZH	Chinese	VA	2,802
Imbir (2016)	PL	Polish	VAD	4,905
Montefinese et al. (2014)	IT	Italian	VAD	1,121
Soares et al. (2012)	PT	Portuguese	VAD	1,034
Moors et al. (2013)	NL	Dutch	VAD	4,299
Sianipar et al. (2016)	ID	Indonesian	VAD	1,490

Table 1: Emotion lexicons used in our experiments (with their bibliographic source, identifier, language they refer to, emotion representation format, and number of lexical entries they contain).

Word	Valence	Arousal	Dominance
sunshine	8.1	5.3	5.4
terrorism	1.6	7.4	2.7
orgasm	8.0	7.2	5.8

Table 2: Three sample entries from Warriner et al. (2013). They use 9-point scales ranging from 1 (most negative/calm/submissive) to 9 (most positive/excited/dominant).

word2vec (with its variants Sgns and Cbow) features an extremely trimmed down neural network (Mikolov et al., 2013). FastText is a derivative of word2vec, also incorporating sub-word character n-grams (Bojanowski et al., 2017). Unlike the former two algorithms which fit word embeddings in a streaming fashion, GloVe trains word vectors directly on a word co-occurrence matrix under the assumption to make more efficient use of word statistics (Pennington et al., 2014). Somewhat similar, SVD$_{\text{PPMI}}$ performs singular value decomposition on top of a point-wise mutual information co-occurrence matrix (Levy et al., 2015).

In order to increase the reproducibility of our experiments, we rely on the following widely used, publicly available embedding models trained on very large corpora (summarized in Table 3): the Sgns model trained on the Google News corpus[2] (Google), the FastText model trained on Common Crawl[3] (Common), as well as the FastText models for a wide range of languages trained on the respective Wikipedias[4] (Wiki). Note that Wiki denotes multiple embedding models with different training and vocabulary sizes (see Grave et al. (2018) for further details). Additionally, we were given the opportunity to reuse the English embedding model from Sedoc et al. (2017) (Giga), a strongly related contribution (see below). Their embeddings were trained on the English Gigaword corpus (Parker et al., 2011).

Word-Level Prediction. One of the early approaches to word *polarity* induction which is still popular today (Köper and Schulte im Walde, 2016) was introduced by Turney and Littman (2003). They compute the polarity of an unseen word based on its point-wise mutual information (PMI) to a set of positive and negative seed words, respectively.

SemEval-2015 Task 10E featured polarity induction on Twitter (Rosenthal et al., 2015). The best system relied on support vector regression (SVR) using a radial base function kernel (Amir et al., 2015). They employ the embedding vector of the target word as features. The results of their SVR-based system were beaten by the Densifier algorithm (Rothe et al., 2016). Densifier learns an orthogonal transformation of an embedding space into a subspace of strongly reduced dimensionality.

Hamilton et al. (2016) developed SentProp, a graph-based, semi-supervised learning algorithm which builds up a word graph, where vertices correspond to words (of known as well as unknown polarity) and edge weights correspond to the similarity between them. The polarity information is then propagated through the graph, thus computing scores for unlabeled nodes. According to their evaluation, Densifier seems to be superior overall, yet SentProp produces competitive results

[2] `https://code.google.com/archive/p/word2vec/`

[3] `https://fasttext.cc/docs/en/english-vectors.html`

[4] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

ID	Language	Method	Corpus	# Tokens	# Types	# Dimensions
Google	English	SGNS	Google News	1×10^{11}	3×10^{6}	300
Common	English	FastText	Common Crawl	6×10^{11}	2×10^{6}	300
Giga	English	CBOW	Gigawords	4×10^{9}	2×10^{6}	300
Wiki	all	FastText	Wikipeda	—	—	300

Table 3: Embedding models used for our experiments with identifier, language, embedding algorithm, training corpus, its size in the number of tokens, size of the vocabulary (types) of the resulting embedding model and its dimensionality.

only when the seed lexicon or the corpus the word embeddings are trained on is very small.[5]

For word *emotion* induction, a very similar approach to SentProp has been proposed by Wang et al. (2016a). They also propagate affective information (Valence and Arousal, in this case) through a word graph with similarity weighted edges.

Sedoc et al. (2017) recently proposed an approach based on signed spectral clustering where a word graph is constructed not only based on word similarity but also on the considered affective information (again, Valence and Arousal). The emotion value of a target word is then computed based on the seed words in its cluster. They report to outperform the results from Wang et al. (2016a).

Contrary to the trend to graph-based methods, the best system of the IALP 2016 Shared Task on Chinese word emotion induction (Yu et al., 2016b) employed a simple feed-forward neural network (FFNN) with one hidden layer in combination with boosting (Du and Zhang, 2016).

Another very recent contribution which advocates a supervised set-up was published by Li et al. (2017). They propose ridge regression, again using word embeddings as features. Even with this simple approach, they report to outperform many of the above methods in the VAD prediction task.[6]

Sentence-Level and Text-Level Prediction. Different from the word-level prediction task (the one we focus on in this contribution), the determination of emotion values for higher-level linguistic units (especially sentences and texts) is also heavily investigated. For this problem, DL approaches are meanwhile fully established as the method of choice (Wang et al., 2016b; Abdul-Mageed and Ungar, 2017; Felbo et al., 2017; Mohammad and Bravo-Marquez, 2017).

It is important to note, however, that the methods discussed for these higher-level units cannot easily be transferred to solve the word emotion induction problem. Sentence-level and text-level architectures are either adapted to *sequential* input data (typical for RNN, LSTM, GRNN and related architectures) or *spatially arranged* input data (as with CNN architectures). However, for word embeddings (the default input for word emotion induction) there does not seem to be any meaningful order of their components. Therefore, these more sophisticated DL methods are, for the time being, not applicable for the study at hand.

3 Methods

In this section, we will first introduce various reference methods (two originally polarity-based for which we offer adaptations for VAD prediction) before defining our own neural MTL model and discussing its difference from previous work.

Let $V := \{w_1, w_2, ..., w_m\}$ be our word vocabulary and let $E := \{e_1, e_2, ..., e_m\}$ be a set of embedding vectors such that $e_i \in \mathbb{R}^n$ denotes the n-dimensional vector representation of word w_i. Let $D := \{d_1, d_2, ..., d_l\}$ be a set of emotional dimensions. Our task is to predict the empirically determined emotion vector $emo(w) \in \mathbb{R}^l$ given a word w and the embedding space E.

3.1 Reference Methods

Linear Regression Baseline (LinReg). We propose (multi-variate) linear regression as an obvious baseline for the problem:

$$emo_{LR}(w_k) := We_k + b \tag{1}$$

where W is a matrix, W_{i*} contains the regression coefficients for the i-th affective dimension and b is the vector of bias terms. The model parameters are fitted using ordinary least squares. Technically, we use the `scikit-learn.org` implementation with default parameters.

[5] Personal correspondence with William L. Hamilton; See also README at `https://github.com/williamleif/socialsent`

[6] However, they also report extremely weak performance figures for some of their reference methods.

Ridge Regression (RidgReg). Li et al. (2017) propose ridge regression for word emotion induction. Ridge regression works identically to linear regression during prediction, but introduces L_2 regularization during training. Following the authors, for our implementation, we again use the `scikit-learn` implementation with default parameters.

Turney-Littman Algorithm (TL). As one of the earliest contributions in the field, Turney and Littman (2003) defined a simple PMI-based approach to determine the semantic polarity SP_{TL} of a word w:

$$SP_{TL}(w) := \sum_{s \in seeds^+} pmi(w,s) - \sum_{s \in seeds^-} pmi(w,s) \tag{2}$$

where $seeds^+$ and $seeds^-$ are sets of positive and negative seed words, respectively. Since this algorithm is still popular today (Köper and Schulte im Walde, 2016), we here provide a novel modification for adapting this originally polarity-based approach to word emotion induction with vectorial seed and output values.

First, we replace PMI-based association of seed and target word w and s by their similarity sim based on their word embeddings e_w and e_s:

$$sim(w,s) := max(0, \frac{e_w \cdot e_s}{||e_w|| \times ||e_s||}) \tag{3}$$

$$emo(w) := \sum_{s \in seeds^+} sim(w,s) - \sum_{s \in seeds^-} sim(w,s) \tag{4}$$

Although this step is technically not required for the adaptation, it renders the TL algorithm more comparable to the other approaches evaluated in Section 4 besides from most likely increasing performance. Equation (4) can be rewritten as

$$emo(w) := \sum_{s \in seeds} sim(w,s) \times emo(s) \tag{5}$$

where $seeds := seeds^+ \cup seeds^-$ and $emo(s)$ maps to 1, if $s \in seeds^+$, and -1, if $s \in seeds^-$.

Equation (5) can be trivially adapted to an n-dimensional emotion format by redefining $emo(s)$ such that it maps to a vector from $\mathbb{R}^n$ instead of $\{-1, 1\}$. Our last step is to introduce a normalization term such that $emo(w)_{TL}$ lies within the range of the seed lexicon.

$$emo_{TL}(w) := \frac{\sum_{s \in seeds} sim(w,s) \times emo(s)}{\sum_{s \in seeds} sim(w,s)} \tag{6}$$

As can be seen from Equation (6), for the more general case of n-dimensional emotion prediction, the Turney-Littman algorithm naturally translates into a weighted average where the seed emotion values are weighted according to the similarity to the target item.

Densifier. Rothe et al. (2016) train an orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ (n being the dimensionality of the word embeddings) such that applying Q to an embedding vector e_i concentrates all the polarity information in its first dimension such that the polarity of a word w_i can be computed as

$$SP_{\text{DENSIFIER}}(w_i) := pQe_i \tag{7}$$

where $p = (1, 0, 0, ..., 0)^T \in \mathbb{R}^{1 \times n}$.

For fitting Q, the seeds are arranged into pairs of equal polarity (the set $pairs^=$) and those of opposing polarity ($pairs^{\neq}$). A good fit for Q will minimize the distance within the former and maximize the distance within the latter which can be expressed by the following two training objectives:

$$\underset{Q}{\text{argmin}} \sum_{(w_i,w_j) \in pairs^=} |pQ(e_i - e_j)| \tag{8}$$

$$\underset{Q}{\text{argmax}} \sum_{(w_i,w_j) \in pairs^{\neq}} |pQ(e_i - e_j)| \tag{9}$$

The objectives described in the expressions (8) and (9) are combined into a single loss function (using a weighting factor $\alpha \in [0, 1]$) which is then minimized using stochastic gradient descent (SGD).

To adapt this algorithm to dimensional emotion formats, we construct a positive seed set, $seeds_v^+$, and a negative seed set, $seeds_v^-$, for each emotion dimension $v \in D$. Let M_v be the mean value of all the entries of the training lexicon for the affective dimension v. Let SD_v be the respective standard deviation and $\beta \in \mathbb{R}$, $\beta \geq 0$. Then all entries greater than $M_v + \beta SD_v$ are assigned to $seeds_v^+$ and those less than $M_v - \beta SD_v$ are assigned to $seeds_v^-$. Q is fitted individually for each emotion dimension v.

Training was performed according to the original paper with the exception that (following Hamilton et al. (2016)) we did not apply the proposed re-orthogonalization after each training step, since we did not find any evidence that this procedure actually results in improved performance. The hyperparameters α and β were set to .7 and .5 (respectively) for all experiments based on a pilot study. Since the original implementation is not accessible, we devised our own using `tensorflow.org`.

Boosted Neural Networks (ensembleNN). Du and Zhang (2016) propose simple FFNNs in combination with a boosting algorithm. An FFNN consists of an *input* or *embedding layer* with activation $a^{(0)} \in \mathbb{R}^n$ which is equal to the embedding vector e_k when predicting the emotion of a word w_k. The input layer is followed by multiple hidden layers with activation

$$a^{(l+1)} := \sigma(W^{(l+1)}a^{(l)} + b^{(l+1)}) \qquad (10)$$

where $W^{(l+1)}$ and $b^{(l+1)}$ are the weights and biases for layer $l+1$ and σ is a nonlinear activation function. Since we treat emotion prediction as a regression problem, the activation on the output layer a^{out} (where out is the number of non-input layers in the network) is computed as the affine transformation

$$a^{(out)} := W^{(out)}a^{(out-1)} + b^{(out)} \qquad (11)$$

Boosting is a general machine learning technique where several weak estimators are combined to form a strong estimator. The authors used FFNNs with a single hidden layer of 100 units and rectified linear unit (ReLU) activation. The boosting algorithm AdaBoost.R2 (Drucker, 1997) was used to train the ensemble (one per affective dimension). Our re-implementation copies their technical set-up[7] exactly using `scikit-learn`.

3.2 Multi-Task Learning Neural Network

The approaches introduced in Section 3.1 and Section 2 vary largely in their methodological foundations, i.e., they comprise semi-supervised and supervised machine learning techniques—both statistical and neural ones. Yet, they all have in common that they treat the prediction of the different emotional dimensions *as separate* tasks. That is, they fit one individual model per VAD dimension without sharing parameters between them.

In contradistinction, the key feature of our approach is that we fit a single FFNN model to predict *all VAD dimensions jointly*, thus applying multi-task learning to word emotion induction. Hence, we treat the prediction of Valence, Arousal and Dominance as three independent tasks. Our multi-task learning neural network (MTLNN) (depicted in Figure 2) has an output layer of three units such that each output unit represents one of the VAD dimensions. However, the activation in our two hidden layers (of 256 and 128 units, respectively) is *shared* across all VAD dimensions, and so are the associated weights and biases.

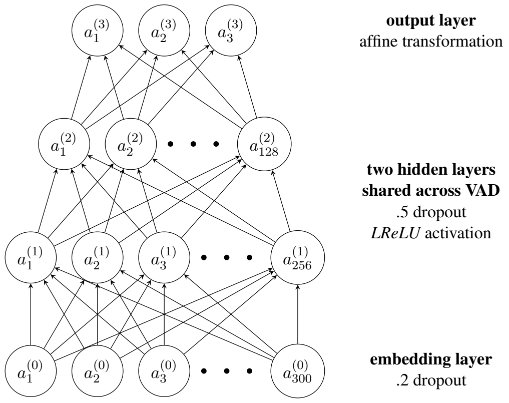


Figure 2: MTL architecture for VAD prediction.

Thus, while we train our MTLNN model it is forced to learn intermediate representations of the input which are generally informative for all VAD dimensions. This serves as a form of regularization, since it becomes less likely for our model to fit the noise in the training set as noise patterns may vary across emotional dimensions. Simultaneously, this has an effect similar to an increase of the training size, since each sample now leads to additional error signals during backpropagation. Intuitively, both properties seem extremely useful for relatively small-sized emotion lexicons (see Section 4 for empirical evidence).

The remaining specifications of our model are as follows. We use *leaky* ReLU activation (LReLU) as nonlinearity (Maas et al., 2013).

$$LReLU(z_i) := max(\gamma z_i, z_i) \qquad (12)$$

with $\gamma := .01$ for our experiments. For regularization, dropout (Srivastava et al., 2014) is applied during training with a probability of .2 on the embedding layer and .5 on the hidden layers. We train for $15,000$ iterations (well beyond convergence on each data set we use) with the ADAM optimizer (Kingma and Ba, 2015) of .001 base learning rate,

[7] Original settings available at `https://github.com/StevenLOL/ialp2016_Shared_Task`

batch size of 128 and Mean-Squared-Error loss. The weights are randomly initialized (drawn from a normal distribution with a standard deviation .001) and biases are uniformly initialized as .01. `Tensorflow` is used for implementation.

4 Results

In this section, we first validate our assumption that MTL is superior to single-task learning for word emotion induction. Next, we compare our proposed MTLNN model in a large-scale evaluation experiment.

Performance figures will be measured as Pearson correlation (r) between our automatically predicted values and human gold ratings. The Pearson correlation between two data series $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_n$ takes values between $+1$ (perfect positive correlation) and -1 (perfect negative correlation) and is computed as

$$r_{xy} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{13}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values for X and Y, respectively.

4.1 Single-Task vs. Multi-Task Learning

The main hypothesis of this contribution is that an MTL set-up is superior to single-task learning for word emotion induction. Before proceeding to the large-scale evaluation of our proposed model, we will first examine this aspect of our work.

For this, we use the following experimental set-up: We will compare the MTLNN model against its single-task learning counterpart (SepNN). SepNN simultaneously trains three separate neural networks where only the input layer, yet no parameters of the intermediate layers are shared across the models. Each of the separate networks is identical to MTLNN (same layers, dropout, initialization, etc.), yet has only one output neuron, thus modeling only one of the three affective VAD dimensions. SepNN is equivalent to fitting our proposed model (but with only one output unit) to the different VAD dimensions individually, one after the other. Yet, training these separate networks simultaneously (not jointly!) makes both approaches, MTLNN and SepNN, easier to compare.

We will run MTLNN against SepNN on the EN and the EN+ data set (the former is very small, the latter relatively large; see Table 1) using the following set-up: for each gold lexicon and model, we randomly split the data $9/1$ and train for $15,000$ iterations on the larger split (the same number of steps is used for the main experiment). After each one-thousand iterations step, model performance is tested on the held-out data. This process will be repeated 20 times and the performance figures at each one-thousand iterations step will be averaged. In a final step, we will average the results for each of the three emotional dimensions and only plot this average value. The results of this experiment are depicted in Figure 3.

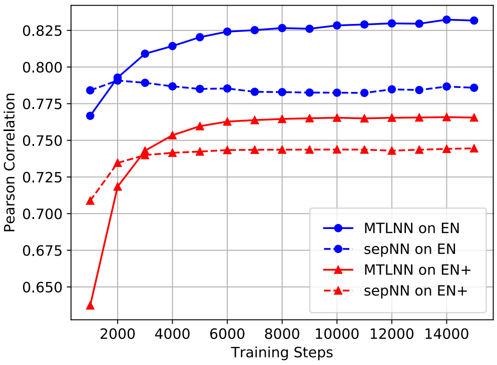


Figure 3: Performance of our proposed MTLNN model vs. its single-task learning counterpart SepNN against training steps.

First of all, each combination of model and data set displays a satisfactory performance of at least $r \approx .75$ after 15,000 steps compared to previous work (see below). Overall, performance is higher for the smaller EN lexicon. Although counterintuitive (since smaller lexicons lead to fewer training samples), this finding is consistent with prior work (Sedoc et al., 2017; Li et al., 2017) and is probably related to the fact that smaller lexicons usually comprise a larger portion of strongly emotion-bearing words. In contrast, larger lexicons add more neutral words which tend to be harder to predict in terms of correlation.

As hypothesized, the MTLNN model does indeed outperform the single task model on both data sets. Our data also suggest that the gain from the MTL approach is larger on smaller data sets (again in concordance with our expectations). Figure 3 reveals that this might be due to the regularizing effect of MTL, since the SepNN model shows signs of overfitting on the EN data set. Yet, even

Language	Data	Embeddings	LinReg	RidgReg	TL	Densifier	ensembleNN	MTLNN
English	EN+	Google	0.696	0.696	0.631	0.622	0.728	**0.739*****
English	EN+	Common	0.719	0.719	0.659	0.652	0.762	**0.767*****
English	EN+	Wiki	0.666	0.666	0.591	0.584	0.706	**0.712*****
English	EN	Google	0.717	0.732	0.723	0.712	0.688	**0.810*****
English	EN	Common	0.731	0.741	0.741	0.726	0.717	**0.824*****
English	EN	Wiki	0.656	0.667	0.674	0.665	0.681	**0.777*****
Spanish	ES	Wiki	0.698	0.709	0.704	0.690	0.700	**0.804*****
Spanish	ES+	Wiki	0.693	0.694	0.603	0.598	0.766	**0.778*****
German	DE	Wiki	0.709	0.719	0.714	0.710	0.700	**0.801*****
Chinese	ZH	Wiki	0.716	0.717	0.586	0.599	0.737	**0.744****
Polish	PL	Wiki	0.650	0.650	0.577	0.553	0.687	**0.712*****
Italian	IT	Wiki	0.656	0.665	0.672	0.659	0.630	**0.751*****
Portuguese	PT	Wiki	0.673	0.684	0.685	0.678	0.672	**0.768*****
Dutch	NL	Wiki	0.651	0.652	0.559	0.532	0.704	**0.730*****
Indonesian	ID	Wiki	0.581	0.586	0.581	0.576	0.575	**0.660*****
Average			0.638	0.659	0.611	0.605	0.676	**0.728*****

Table 4: Results of our main experiment in averaged Pearson correlation; best result per condition (in rows) in bold, second best result underlined; significant difference (paired two-tailed t-test) over the second best system marked with "*", "**", or "***" for $p < .05$, .01, or .001, respectively.

when the separate model does not overfit (as on the EN+ lexicon), MTLNN reveals better results.

Although SepNN needs fewer *training steps* before convergence, the MTLNN model trains much faster, thus still converging faster in terms of *runtime* (about a minute on a middle-class GPU). This is because MTLNN has only about a third as many parameters as the separate model SepNN.

4.2 Comparison against Reference Methods

We combined each of the selected lexicon data sets (Table 1) with each of the applicable publicly available embedding models (Section 2; the embedding model provided by Sedoc et al. (2017) will be used separately) for a total of 15 conditions, i.e, the rows in Table 4.

For each of these conditions, we performed a 10-fold cross-validation (CV) for each of the 6 methods presented in Section 3 such that each method is presented with the identical data splits.[8] For each condition, algorithm, and VA(D) dimension, we compute the Pearson correlation r between gold ratings and predictions. For conciseness, we present only the average correlation over the respective affective dimensions in Table 4 (Valence and Arousal for ES+ and ZH, VAD for the others). Note that the methods we compare ourselves against comprise the current state-of-the art in both polarity and emotion induction (as described in Section 2).

As can be seen, our proposed MTLNN model outperforms all other approaches in each of the 15 conditions. Regarding the average over all affective dimensions and conditions, it outperforms the second best system, ensembleNN, by more than 5%-points. In line with our results from Section 4.1, those improvements are especially pronounced on smaller data sets containing one up to two thousand entries (EN, ES, IT, PT, ID) with close to 10%-points improvement over the respective second-best system.

Concerning the relative ordering of the affective dimensions, in line with former studies (Sedoc et al., 2017; Li et al., 2017), the performance figures for the Valence dimension are usually much higher than for Arousal and Dominance. Using MTLNN, for many conditions, we see the pattern that Valence is about 10%-points above the VAD average, Arousal being 10%-points below and Dominance being roughly equal to the average over VAD (this applies, e.g., to EN, EN+ and IT). On other data sets (e.g., PL, NL and ID), the ordering between Arousal and Dominance is less clear though Valence still stands out with the best results. We observe the same general pattern for the reference methods, as well.

Concerning the comparison to Sedoc et al. (2017), arguably one of most related contributions, they report a performance of $r = .768$ for Valence and .582 for Arousal on the EN+ data set in a 10-fold CV using their own embeddings. In contrast, MTLNN using the Common model achieves $r = .870$ and .674 in the same set-up—about 10%-

[8] This procedure constitutes a more direct comparison than using different splits for each method and allows using *paired* t-tests.

	Valence	Arousal	Dominance
MTLNN EN	.918	.730	.825
MTLNN EN+	.870	.674	.758
ISR EN $\sim$ EN+	.953	.759	.795
SHR EN+	.914	.689	.770

Table 5: Comparison of the MTLNN model against inter-study reliability (ISR) between the EN and the EN+ data set and split-half reliability (SHR) of the EN+ data set (in Pearson correlation).

points better on both dimensions. However, the COMMON model was trained on much more data than the embeddings Sedoc et al. (2017) use. For the most direct comparison, we also repeated this experiment using *their* embedding model (GIGA). We find that MTLNN still clearly outperforms their results with $r = .814$ for Valence and .607 for Arousal.[9]

MTLNN achieves also very strong results in direct comparison to human performance (see Table 5). Warriner et al. (2013) (who created EN+) report an inter-study reliability (ISR; i.e., the correlation of the aggregated ratings from two different studies) between the EN and the EN+ lexicon of $r = .953$, .759 and .795 for VAD, respectively. Since EN is a subset of EN+, we can compare these performance figures against our own results on the EN data set where we achieved $r = .918$, .730 and .825, respectively. Thus, our proposed method did actually outperform human reliability for Dominance and is competitive for Valence and Arousal, as well.

This general observation is also backed up by split-half reliability data (SHR; i.e., when randomly splitting all individual ratings in two groups and averaging the ratings within each group, how strong is the correlation between these averaged ratings?). For the EN+ data set, Warriner et al. (2013) report an SHR of $r = .914$, .689 and .770 for VAD, respectively. Again, our MTLNN model performs very competitive with $r = .870$, .674 and .758, respectively using the COMMON embeddings.

5 Conclusion

In this paper, we propose multi-task learning (MTL) as a simple, yet surprisingly efficient method to improve the performance and, at the same time, to deal with existing data limitations in word emotion induction—the task to predict a complex emotion score for an individual word. We validated our claim that MTL is superior to single-task learning by achieving better results with our proposed method in performance as well as training time compared to its single-task counterpart. We performed an extensive evaluation of our model on 9 typologically diverse languages, using different kinds of word embedding models for a total 15 conditions. Comparing our approach to state-of-the-art methods from word polarity and word emotion induction, our model turns out to be superior in each condition, thus setting a novel state-of-the-art performance for both polarity *and* emotion induction. Moreover, our results are even competitive to human annotation reliability in terms of inter-study as well as split-half reliability. Since this contribution was restricted to the VAD format of emotion representation, in future work we will examine whether MTL yields similar gains for other representational schemes, as well.

Acknowledgments

We would like to thank the Positive Psychology Center, University of Pennsylvania for providing us with the embedding model used in Sedoc et al. (2017), Johannes Hellrich, JULIE Lab, for insightful discussions, and the reviewers for their valuable comments.

References

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. EMONET: Fine-grained emotion detection with gated recurrent neural networks. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, Canada, July 30 - August 4, 2017, volume 1: Long Papers, pages 718–728.

Silvio Amir, Ramón F. Astudillo, Wang Ling, Bruno Martins, Mário J. Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale Twitter sentiment lexicon induction. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*. Denver, Colorado, USA, June 4-5, 2015, pages 613–618.

Jonathan Baxter. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning* 28(1):7–39.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5(1):135–146.

[9] We also clearly outperform their results for the NL and ES+ data sets. For these cases, our embedding models were similar in training size.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*. The Hague, Netherlands, August 29 - September 2, 2016, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 1114–1122.

Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, April 3–7, 2017, volume 2: Short Papers, pages 578–585.

Sven Buechel and Udo Hahn. 2018. Representation mapping: A novel approach to generate high-quality multi-lingual emotion lexicons. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*. Miyazaki, Japan, May 7–12, 2018.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence* 29(3):527–543.

Rich Caruana. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, August 7-12, 2016, volume 2: Short Papers, pages 789–799.

Harris Drucker. 1997. Improving regressors using boosting techniques. In *ICML '97 — Proceedings of the 14th International Conference on Machine Learning*. Nashville, Tennessee, USA, July 8-12, 1997, pages 107–115.

Steven Du and Xi Zhang. 2016. Aicyber's system for IALP 2016 Shared Task: Character-enhanced word vectors and boosted neural networks. In *IALP 2016 — Proceedings of the [20th] 2016 International Conference on Asian Language Processing*. Tainan, Taiwan, November 21-23, 2016, pages 161–163.

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

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, September 9-11, 2017, pages 1615–1625.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*. Miyazaki, Japan, May 7–12, 2018.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Daniel Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, USA, November 1-5, 2016, pages 595–605.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997 — Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, July 7-12, 1997, pages 174–181.

Kamil K. Imbir. 2016. Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology* 7:#1081.

Diederik Kingma and Jimmy Ba. 2015. ADAM: A method for stochastic optimization. In *ICLR 2015 — Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California, USA, May 7-9, 2015.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*. Portorož, Slovenia, 23-28 May 2016, pages 2595–2598.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing* 8(4):443–456.

PengFei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, Canada, July 30 - August 4, 2017, volume 1: Long Papers, pages 1–10.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *NAACL-HLT 2015 — Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, USA, May 31 - June 5, 2015, pages 912–921.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the Workshop on Deep Learning for Audio, Speech and Language Processing @ ICML 2013*. Atlanta, Georgia, USA, 16 June 2013.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, USA, December 5-10, 2013, pages 3111–3119.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *WASSA 2017 — Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2017*. Copenhagen, Denmark, September 8, 2017, pages 34–49.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods* 46(3):887–903.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbært. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods* 45(1):169–177.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. Technical Report LDC2011T07, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA. `https://catalog.ldc.upenn.edu/LDC2011T07`.

Hao Peng, Sam Thomson, and Noah A. Smith. 2017. Deep multitask learning for semantic dependency parsing. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, Canada, July 30 - August 4, 2017, volume 1: Long Papers, pages 2037–2048.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GLOVE: Global vectors for word representation. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, October 25-29, 2014, pages 1532–1543.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods* 39(3):600–605.

Sara Rosenthal, Preslav I. Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval 2015 Task 10: Sentiment analysis in Twitter. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*. Denver, Colorado, USA, June 4-5, 2015, pages 451–463.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, USA, June 12-17, 2016, pages 767–777.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161–1178.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11(3):273–294.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. 2014. ANGST: Affective Norms for German Sentiment Terms, derived from the Affective Norms for English Words. *Behavior Research Methods* 46(4):1108–1118.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle H. Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, April 3-7, 2017, volume 2: Short Papers, pages 564–571.

Hendra Setiawan, Zhongqiang Huang, Jacob Devlin, Thomas Lamar, Rabih Zbib, Richard M. Schwartz, and John Makhoul. 2015. Statistical machine translation features with multitask tensor networks. In *ACL-IJCNLP 2015 — Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics & 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Beijing, China, July 26-31, 2015, volume 1: Long Papers, pages 31–41.

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology* 7:#1907.

Ana Paula Soares, Montserrat Comesaña, Ana P Pinheiro, Alberto Simões, and Carla Sofia Frade. 2012. The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods* 44(1):256–269.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Berlin, Germany, August 7–12, 2016, volume 2: Short Papers, pages 231–235.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez-Sánchez, and Marc Brysbært. 2017. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods* 49(1):111–123.

Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James. 2007. Characterization of the affective norms for English words by discrete emotional categories. *Behavior Research Methods* 39(4):1020–1024.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* 21(4):315–346.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016a. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(11):1957–1968.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016b. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, August 7-12, 2016, volume 2: Short Papers, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4):1191–1207.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016a. Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, USA, June 12-17, 2016, pages 540–545.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016b. Overview of the IALP 2016 Shared Task on dimensional sentiment analysis for Chinese words. In *IALP 2016 — Proceedings of the [20th] 2016 International Conference on Asian Language Processing*. Tainan, Taiwan, November 21-23, 2016, pages 156–160.

11 Emotion Representation Mapping Mostly Performs on Human Level

Reference

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904.

Author Contributions

Udo Hahn performed supervision and project administration. I performed methodology, model, and software development, experimental design and execution of experiments, as well as data analysis and visualization. Conception and writing were performed jointly by both authors.

Emotion Representation Mapping for Automatic Lexicon Construction (Mostly) Performs on Human Level

Sven Buechel & Udo Hahn
{sven.buechel|udo.hahn}@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
http://www.julielab.de

Abstract

Emotion Representation Mapping (ERM) has the goal to convert existing emotion ratings from one representation format into another one, e.g., mapping Valence-Arousal-Dominance annotations for words or sentences into Ekman's Basic Emotions and vice versa. ERM can thus not only be considered as an alternative to Word Emotion Induction (WEI) techniques for automatic emotion lexicon construction but may also help mitigate problems that come from the proliferation of emotion representation formats in recent years. We propose a new neural network approach to ERM that not only outperforms the previous state-of-the-art. Equally important, we present a refined evaluation methodology and gather strong evidence that our model yields results which are (almost) as reliable as human annotations, even in cross-lingual settings. Based on these results we generate new emotion ratings for 13 typologically diverse languages and claim that they have near-gold quality, at least.

1 Introduction

From its inception, researchers in the field of sentiment analysis aimed at predicting the affective state that is typically associated with a given word based on a list of linguistic features, a problem referred to as *word emotion induction* (WEI) (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003). Early research activities have focused on *semantic polarity* (the positiveness or negativeness of a feeling) for quite a long time. But more recently this focus on binary representations has been replaced by more expressive *emotion representation formats* such as Basic Emotions or Valence-Arousal-Dominance. In the meantime, WEI has become an active area of research, regularly featured in shared tasks (Rosenthal et al., 2015; Yu et al., 2016b). Based on these achievements, WEI techniques have become a natural methodological choice for the automatic construction of emotion lexicons (Köper and Schulte im Walde, 2016; Shaikh et al., 2016).

Yet, only very recently, a radically different approach to automatic emotion lexicon construction has been proposed. Instead of relying on linguistic features (such as similarity with seed words or word embeddings), the goal of *emotion representation mapping* (ERM) is to derive new emotional word ratings *in one format* based on known ratings of the same words *in another format* (Buechel and Hahn, 2017a). For example, ERM could use empirically gathered ratings for Basic Emotions and convert them into a Valence-Arousal-Dominance representation scheme, with greater precision than currently achievable by WEI algorithms. As a much appreciated side effect, one of the promises of ERM is to make otherwise incompatible resources (lexicons or annotated corpora, as well as tools) compatible, and incomparable systems comparable. Thus, this approach has the potential to mitigate some of the negative effects that arise from not having a community-wide standard for emotion annotation and representation (Calvo and Mac Kim, 2013; Buechel and Hahn, 2018a).

We here want to contribute to this endeavor by providing a large-scale evaluation of previously proposed ERM approaches for four typologically diverse languages and report evidence that ERM clearly

Proceedings of the 27th International Conference on Computational Linguistics, pages 2892–2904
Santa Fe, New Mexico, USA, August 20-26, 2018.

outperforms current state-of-the-art WEI algorithms. Furthermore, we present our own deep learning model which performs even better against all competitors. Most importantly, however, we propose a new methodology for comparing the reliability of ERM against human annotation reliability, a major shortcoming of previous work. As a result, we find that our proposed model performs competitive to a reasonably large group of human raters, *even in cross-lingual settings*. Based on this evidence, we automatically construct emotion lexicons for 13 languages and claim that they have (near) gold quality. These lexicons as well as our experimental code base and results are publically available.[1]

2 Related Work

Psychological Models of Emotion. Models of emotion typically fall into two main groups, namely *discrete* (or *categorical*) and *dimensional* ones (Stevenson et al., 2007; Calvo and Mac Kim, 2013). Discrete models are built around particular sets of emotional categories deemed fundamental and universal. Ekman (1992), for instance, identifies six *Basic Emotions* (Joy, Anger, Sadness, Fear, Disgust and Surprise). In contrast, dimensional models consider emotions to be composed out of several influencing factors (mainly two or three). These are often referred to as *Valence* (corresponding to the concept of polarity), *Arousal* (a calm–excited scale), and *Dominance* (perceived degree of control over a (social) situation)—the VAD model. The last dimension, Dominance, is quite often omitted, thus constituting the VA model. For convenience, both will be jointly referred to as VA(D). An illustration of VAD and its relationship to Basic Emotions is given in Figure 1.

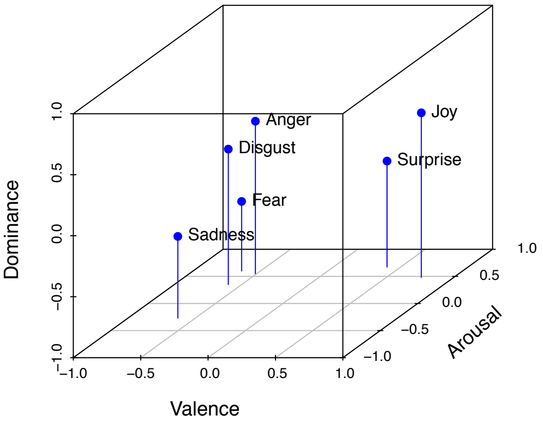


Figure 1: Affective space spanned by the Valence-Arousal-Dominance (VAD) model, together with the position of six Basic Emotions. Adapted from Buechel and Hahn (2016).

Lexical Data Sets. In contradistinction to NLP where many different representation formats for emotions are being used, lexical resources originating from psychology labs almost exclusively subscribe either to VA(D) or Basic Emotions models (typically omitting Surprise; the BE5 format). Over the years, a considerable number of resources built on these premises have emerged from psychological research for various languages.[2] In more detail, these lexical ratings have been gathered via questionnaire studies by collecting individual ratings from a large number of subjects for each lexical item under consideration (typically between 20 to 30 individual ratings per item). These individual assessments are then averaged to yield aggregated scores on which we base our experiments. The emotion values we deal with must thus be understood as an average emotional reaction when presenting a lexical stimulus to a group of human judges.

In this paper, we restrict ourselves to the VA(D) and BE5 format. Following the conventions of the emotion lexicons used in our experiments (Table 1), each VA(D) dimension receives a value from the interval $[1, 9]$ where '1' means "most negative/calm/submissive", '9' means "most positive/excited/dominant" and '5' means "neutral". Conversely, values for BE5 categories range in the interval $[1, 5]$ where '1' means "absence" and '5' means "most extreme" expression of the respective emotion.[3] Consequently, the VA(D) and BE5 formats are conceptually different from one another insofar as VA(D) dimensions are bi-polar, whereas BE5 categories are uni-polar.

[1] `https://github.com/JULIELab/EmoMap`

[2] See, e.g., Tables 1 and 6. An enhanced list of these and similar data sets is provided in Buechel and Hahn (2018a).

[3] Although these intervals are fairly well established conventions, in some data sets different rating scales were used, nevertheless. In these cases, we linearly transformed the ratings so that they match the defined intervals.

Abbrev.	VA(D)	BE5	Dom?	Overlap
en_1	Bradley and Lang (1999)	Stevenson et al. (2007)	✓	1,028
en_2	Warriner et al. (2013)	Stevenson et al. (2007)	✓	1,027
es_1	Redondo et al. (2007)	Ferré et al. (2017)	✓	1,012
es_2	Hinojosa et al. (2016b)	Hinojosa et al. (2016a)	✓	875
es_3	Stadthagen-Gonzalez et al. (2017b)	Stadthagen-González et al. (2017a)	✗	10,491
de_1	Võ et al. (2009)	Briesemeister et al. (2011)	✗	1,958
pl_1	Riegel et al. (2015)	Wierzba et al. (2015)	✗	2,902
pl_2	Imbir (2016)	Wierzba et al. (2015)	✓	1,272

Table 1: Data sets used in our experiments; with abbreviation (including language code according to ISO 639-1), the bibliographic sources of the VA(D) and BE5 ratings, information on whether Dominance is included and the number of overlapping entries.

Word Emotion Induction. Automatically constructing such word-level emotion data sets has been a focus of NLP-based sentiment analysis studies from the beginning. In fact, the problem to automatically predict polarity or emotion scores for a given word based on some linguistic features—often referred to as Word Emotion Induction (WEI)—is already dealt with in the seminal work of Hatzivassiloglou and McKeown (1997). At first, the features taken into account were typically derived from co-occurrence or terminology-based similarity with a small set of *seed word* with known emotional scores (Turney and Littman, 2003; Esuli and Sebastiani, 2005). Nowadays, these features are almost completely replaced by *word embeddings*, i.e., dense, low-dimensional vector representations of words that are trained on large volumes of raw text in an unsupervised manner. WORD2VEC (Mikolov et al., 2013), GLOVE (Pennington et al., 2014) and FASTTEXT (Bojanowski et al., 2017) are among today's most popular algorithms for generating embeddings.

WEI algorithms constitute a natural baseline for ERM because, first, they produce the same output (emotion ratings for words according to some emotion representation format), yet their predictions are based on expressively weaker features (word embeddings instead of emotion ratings for the same word but in another format), thus constituting a harder task. Second, they form the currently prevailing paradigm for the automatic construction of emotion lexicons (Köper and Schulte im Walde, 2016; Shaikh et al., 2016), a problem for which ERM offers a promising alternative.

Emotion Representation Mapping. In contrast to WEI, ERM is based on the condition that the pairs of data sets in Table 1 are complementary in the sense that, when combining these lexicons, a subset of their entries are then encoded in *both* emotion formats, i.e., VA(D) *and* BE5. This condition is illustrated for three lexical items in Table 2.

Although such complementary data sets have been available for quite some time, ERM has only recently been introduced to NLP by Buechel and Hahn (2016) in order to compare a newly proposed VAD-based prediction system against previously established results on Basic Emotion gold standards. In a follow-up study, Buechel and Hahn (2017b) devised EMOBANK, a VAD-annotated corpus which, in part, also bears BE5 ratings on the *sentence* level. They found that both kinds of annotation were highly predictive for each other using a k-Nearest-Neighbor approach. In later studies, they examined the potential of ERM as a substitute for manual annotation of *lexical* items, also in cross-lingual settings (Buechel and Hahn, 2017a; Buechel and Hahn, 2018a). Although their evaluation was limited in expressiveness, they already found evidence that ERM may be comparable to human performance in terms of the quality of the resulting ratings.

Word	V	A	D	J	A	S	F	D
sunshine	8.1	5.3	5.4	4.3	1.2	1.3	1.3	1.2
terrorism	1.6	7.4	2.7	1.1	3.0	3.4	4.1	2.5
orgasm	8.0	7.2	5.8	4.3	1.3	1.3	1.4	1.2

Table 2: Three lexical items and their emotion values in VAD (second column group) and BE5 (third column group) format. VAD scores are taken from Warriner et al. (2013), BE5 scores were automatically derived (see Section 4.4).

Similar work has, to the best of our knowledge, only been done in the psychology domain. However, related work from this area does not target the goal of predictive modeling (Stevenson et al., 2007; Pinheiro et al., 2017). In both contributions, linear regression models were fitted to predict VAD dimensions given BE5 categories and vice versa. Yet, this was mainly done to inspect the respective slope-coefficients as an indicator of the relationship of dimensions and categories. Thus, the overall goodness of the fit was *not* in the center of interest and was not even reported by Stevenson et al. (2007).

3 Methods

Let $L := \{w_1, w_2, ..., w_n\}$ be a set of words. Let s,t denote two distinct *emotion representation formats* such that *both* $emo^s(w_i) \in \mathbb{R}^{|s|}$ and $emo^t(w_i) \in \mathbb{R}^{|t|}$ describe the emotion vector associated with w_i relative to s and t, respectively, where $|s|$, $|t|$ denote the number of variables which each format employs (e.g., 3 for VAD and 5 for BE5). The task we address in this paper is to predict the *target emotion ratings* $T := \{emo^t(w_i)|\ w_i \in L\}$ given the set L and the corresponding *source emotion ratings* $S := \{emo^s(w_i)|\ w_i \in L\}$. Performance will be measured as Pearson correlation r between the predicted values and human gold ratings (one r-value per element of the target representation). In general, the Pearson correlation between two data series $X := x_1, x_2, ..., x_n$ and $Y := y_1, y_2, ..., y_n$ takes values between $+1$ (perfect positive correlation) and -1 (perfect negative correlation) and is computed as

$$r_{xy} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values for X and Y, respectively.

3.1 Reference Methods

The first method against which we will compare our proposed model is linear regression (LR) as used by Stevenson et al. (2007) in their early study. LR predicts an emotion value in the target representation t as the affine transformation

$$emo^t_{\text{LR}}(w_i) := W\ emo^s(w_i) + b \tag{2}$$

where W is a $|t| \times |s|$ matrix and b is a $|t| \times 1$ vector. The model parameters are fitted using ordinary least squares. In contrast, Buechel and Hahn (2017b) proposed the use of k-Nearest-Neighbor Regression (KNN) for ERM. This simple supervised approach predicts the target value as

$$emo^t_{\text{KNN}}(w_i) := \frac{1}{k} \sum_{w_i{}' \in \text{NEAREST}(w_i,k,S)} emo^t(w_i{}') \tag{3}$$

where NEAREST yields the k nearest neighbors of w_i in the training set (determined by the Euclidean distance between the source representations of two words). The k parameter was fixed to 20 based on a pilot study.[4] We used the `scikit-learn.org` implementation for both LR and KNN.

3.2 Proposed Model: A Multi-Task Feed-Forward Neural Network for ERM

Despite the fact that the above set-ups already perform quite well for ERM (see Section 4), both LR and KNN are rather basic types of models lacking deeper sophistication. As a consequence, we here propose the use of Feed-Forward Neural Networks[5] (FFNNs) for ERM which have been shown to be capable of approximating arbitrary functions, in theory at least (Hornik, 1991). In general, an FFNN consists of an *input layer* with activation $a^{(0)} := emo^s(w_i) \in \mathbb{R}^{|s|}$ followed by multiple hidden layers with activation $a^{(l+1)} := \sigma(W^{(l+1)}a^{(l)} + b^{(l+1)})$ where $W^{(l+1)}$ and $b^{(l+1)}$ are the weights and biases for layer $l + 1$ and σ is a nonlinear activation function. Since the emotion formats under scrutiny capture affective states as real-valued vectors, the activation on the output layer a^{out} (where out is the number of non-input layers in the network) is computed as the affine transformation

$$emo^t_{\text{FFNN}}(w_i) := a^{(out)} := W^{(out)}a^{(out-1)} + b^{(out)} \tag{4}$$

[4] In contrast, Buechel and Hahn (2017a) determined k for each lexicon *individually* based on a dev set. Now, we deviate from this approach since it is inapplicable for the cross-lingual lexicon construction presented in Section 4.4.

[5] Note that applying neural architectures currently popular for other NLP tasks is not advisable because of the simplicity of our input data (feature vectors of length 2 to 5). These more complex architectures are instead designed for, e.g., *sequential* data (such as the RNN family) or *spatially arranged* data (such as CNNs).

Consequently, our model differs from the other approaches presented in this section by *sharing* model parameters (weights and biases of the hidden layers) across the different dimensions/categories of the target format with only the last layer having parameters which are uniquely associated to one of the outputs (see Equation 4). This can be considered as a mild form of multi-task learning (Caruana, 1997), a machine learning technique which has been shown to strongly decrease the risk of overfitting (Baxter, 1997) and also speeds up computation by greatly decreasing the number of tunable parameters compared to training individual layers for each affective dimension/category.

The remaining specifications of our model are as follows. We train two-hidden layer FFNNs (both with 128 units), ReLU activation, .2 dropout on the hidden layers (none on the input layer)[6] and Mean-Squared-Error loss. Each model was trained for $10,000$ iterations (well beyond convergence, independently of the size of the training set) using the ADAM optimizer (Kingma and Ba, 2015). `Keras.io` was used for implementation.

3.3 Baseline: Word Emotion Induction

As a natural baseline for ERM, we will use a recent state-of-the-art method for word emotion induction (WEI) by Du and Zhang (2016).[7] They propose Feed-Forward Neural Networks (similar to our proposed model for ERM) in combination with a boosting algorithm. The authors used FFNNs with a single hidden layer of 100 units and ReLU activation. The boosting algorithm ADABOOST.R2 (Drucker, 1997) was used to train the ensemble (one per target variable). We implemented this approach with `scikit-learn` using exactly the same settings as in the original publication.[8] As for the word embeddings this method needs as input, we used the pre-trained FASTTEXT embeddings that Facebook Research makes available for a wide range of languages trained on the respective Wikipedias.[9] This way, we hope to achieve a particularly high level of comparability across languages because, for each of them, embeddings are trained on data from the same domain and of a similar order of magnitude.[10]

3.4 Comparison to Human Reliability

Since common metrics for Inter-Annotator Agreement (IAA), such as Cohen's Kappa, are not applicable for real-valued emotion scores (Carletta, 1996), we will now discuss how to compare our own results against human assessments in order to put their reliability on a safe ground.

One possible point of comparison that has been used in previous work (Buechel and Hahn, 2017a; Buechel and Hahn, 2018a) is *inter-study reliability* (ISR), i.e., the correlation between the ratings of common words in different data sets. However, this procedure comes with a number of downsides. First, the number of pairs of data sets with substantially overlapping entries is rather small since researchers focus mainly on acquiring ratings for *novel* words instead of gathering annotations anew for ones already covered. Thus, employing ISR comparison with human performance is only possible on few data sets. In particular, we are not aware of any pair of data sets with significantly overlapping BE5 ratings. Second, ISR is sensitive to differences in acquisition methodologies (e.g., alternative sets of instructions or rating scales) and may thus vary substantially between different pairs of data sets.

As an alternative, these shortcomings lead us to propose *split-half reliability* (SHR) as a new basis for our comparison. SHR is computed by splitting all individual ratings for each of the items into two groups. These individual ratings are then averaged for both groups and the Pearson correlation between the group averages is computed. The whole processes is repeated (typically 100 times) with random splits before averaging the results from each iteration (Mohammad and Bravo-Marquez, 2017). Thus, an

[6] We found the usual recommendation of .2 on input and .5 on hidden layers (Srivastava et al., 2014) too high given the small number of features in our task (2 to 5).

[7] In our most recent contribution featuring a large-scale evaluation of many current WEI approaches on numerous data sets, we found that among the existing ones the model proposed by Du and Zhang (2016) performs best, only beaten by our own, newly proposed model (Buechel and Hahn, 2018b). Note that even compared to this more advanced approach to WEI, the performance figures we report here for ERM still remain much higher (see Section 4). Hence, the claim of this paper that ERM is superior to WEI, remains valid even despite most recent achievements for the latter task.

[8] Publicly available at: `https://github.com/StevenLOL/ialp2016_Shared_Task`

[9] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

[10] For English, much larger embedding models are publicly available, yet not for the other languages under consideration; cf. Buechel and Hahn (2018b).

important difference between SHR and ISR is that the former is computed on a *single* data set whereas the latter requires two *different* data sets with overlapping items. On the other hand, ISR can be computed on the final ratings alone, whereas SHR requires knowledge of the judgments of the individual raters. Most often, these individual ratings are not distributed. Yet, luckily, SHR values are commonly reported when publishing emotion lexicons (see below).

Still, both SHR and ISR—as well as other popular approaches to reliability estimation for numerical emotion scores, e.g., the leave-one-out approach presented by Strapparava and Mihalcea (2007)—are heavily influenced by the number of participants of a study. For SHR, this is intuitively clear because with enough subjects, both groups should yield reliable estimates of the true population mean ratings, leading to very high correlation values between the groups. As a result, by splitting the number of raters into two groups for the SHR estimate, this technique will on average produce lower correlation values than if the study was repeated with the full number of participants and correlation between the first and second study had been computed (test-retest reliability). To counterbalance this effect, when reporting SHR values, authors often turn to *Spearman-Brown adjustment* (SBA; Vet et al. (2017)), a technique which estimates the reliability r^* of a study if the number of subjects was increased by the factor k:

$$r^* := \frac{k\ r}{1 + (k-1)\ r} \tag{5}$$

were r is the *empirically measured* SHR and k is set to 2 for the use case discussed above (virtually doubling the number participants).

Since some authors of the data sets in Table 1 apply SBA while others do not, the reported SHR values must be normalized to guarantee a consistent evaluation. Going one step further, we can even apply SBA to normalize the reported values with respect to the number of participants in a given study, thus establishing an even more consistent ground for evaluation.

We chose the *normalized number of participants* to be 20, i.e., the adjusted scores (reported in Table 3) estimate the *empirical* SHR values, if the given study was conducted with 20 participants (the average correlation between two randomly assigned groups of 10 raters). Normalization was conducted by applying Equation (5) to the reported values with $k := N^*/N$, if SBA was not already applied, or $k := N^*/(2 \times N)$, if SBA was already applied to the reported values; N being the actual number of participants and $N^* := 20$ being the normalized number of participants.

	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dsg
en_1	—	—	—	—	—	—	—	—
en_2	.914	.689	.770	—	—	—	—	—
es_1	—	—	—	.915	.889	.915	.889	.864
es_2	.839	.730	.730	.915	.915	.915	.889	.889
es_3	.880	.750	—	.754	.786	.818	.802	.739
de_1	—	—	—	—	—	—	—	—
pl_1	.928	.630	—	.884	.802	.821	.821	.802
pl_2	.935	.679	.725	.884	.802	.821	.821	.802

Table 3: Normalized split-half reliabilities for VAD and BE5 for the data sets used in our experiments. "—" indicates that reliability has not been reported.

It is important to note that the decision for $N^* = 20$ is necessarily arbitrary, to some degree, with higher SHR estimates arising from higher values of N^*. However, 20 raters are often used in psychological studies (Warriner et al., 2013; Stadthagen-Gonzalez et al., 2017b), while being way higher than the number of raters typically used in NLP for emotion annotation, both for the word and sentence level (Yu et al., 2016a; Strapparava and Mihalcea, 2007). Thus, we argue that this choice constitutes a rather challenging line of comparison for our system.

Since model performance will be measured in terms of Pearson correlation (see above), the performance figures achieved on the gold data can be compared with the adjusted SHR (also based on correlation). We can interpret cases where the former outperforms the latter as *the model agreeing more with the gold data than two random groups of ten annotators would agree with each other*. Thus, for these cases we say our model achieves *super-human* performance, as it cannot be expected that a well-conducted annotation study leads to more reliable results.

4 Results

4.1 Ablation Experiments on Affective Dimensions and Categories

Previous work has limited itself to data sets comprising all three VAD dimensions with the implicit belief that Dominance provides valuable affective information which is important for ERM. However, since only about half of the data sets developed in psychology labs (and even less provided by NLP groups) actually *do* comprise Dominance, this decision massively decreases the amount of data sets at hand. To resolve this dilemma, the following experiment aims at quantifying the relative importance of the different affective variables of the VAD and the BE5 format.

Our set-up works as follows: For each data set from Table 1 that includes the Dominance dimension, we trained one LR model[11] (Section 3.1) to map VAD to BE5 and another one to map BE5 to VAD ('dim2cat' and 'cat2dim' for short) applying 10-fold cross-validation. The resulting performance measurements were averaged over all data sets.

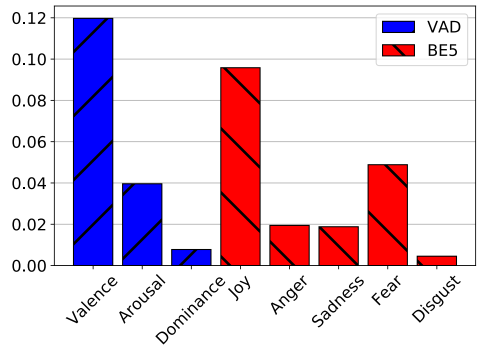


Figure 2: Relative importance of the affective variables of VAD and BE5 for predicting the alternative format, respectively; measured in drop of Pearson r when using all variables vs. omitting the one under scrutiny.

We then repeated this procedure once for each VAD dimension (when mapping dim2cat) and each BE5 category (when mapping cat2dim), omitting one of the dimensions/categories from the source representation in every iteration, thus constituting a kind of ablation experiment. Next, for each of the "incomplete" models, we computed the difference between its performance and the performance of the "complete" model (not lacking any of the variables). Now, we can use this loss of performance as an estimate of the *relative importance* of the respective left-out dimension or category. The results of this experiment are depicted in Figure 2.

As can be seen, regarding VAD, Valence is by far the most important dimension with a performance drop of .12 when ablating it. In turn, Arousal, the second-best dimension only increases performance by .04, whereas Dominance contributes to less than .01 of the performance. Similarly, for Basic Emotions, Joy is the most important category, although BE5 seems to distribute the affective information more equally across its variables (with the exception of Disgust which contributes far less than .01 to the performance).

Since our data suggest that Dominance plays only a minor role within the VAD framework, we will *not* limit our further experiments to data sets including this dimension—as it was done in previous work (Section 2)—but rather include the large variety of bi-representational data sets which leave it out (see Table 1).

4.2 Monolingual Representation Mapping

In this experiment, we compared the performance of the WEI baseline, the LR- and KNN-based reference methods for ERM and our newly proposed FFNN model. For each of these methods and data sets in Table 1, we trained one model to map cat2dim and another one to map dim2cat (for the ERM methods) or to predict VA(D) ratings and BE5 ratings based on word embeddings for the WEI baseline. The whole process was conducted using 10-fold cross-validation where we used identical train/test splits for all methods.[12] The results of this experiment are displayed in Table 4a, only showing the average values over VA(D) and BE5, respectively, but allowing for an easy comparison between the different approaches.

[11] Linear regression was used because it does not comprise any hyperparameters that might heavily influence the outcome of this experiment (thus leading to greater generality of the results).

[12] This procedure constitutes a more direct comparison than using different splits for each method and allows *paired* t-tests.

	cat2dim				dim2cat			
	WEI	LR	KNN	FFNN	WEI	LR	KNN	FFNN
en_1	.685	.841	.840	**.853****	.818	.844	.868	**.877***
en_2	.741	.827	.828	**.843*****	.821	.829	.852	**.858*****
es_1	.709	.856	.855	**.869*****	.775	.804	.849	**.853**
es_2	.600	.823	.828	**.844*****	.797	.863	.882	**.889***
es_3	.713	.799	.796	**.804*****	.743	.776	.820	**.826*****
de_1	.758	.819	.827	**.837****	.701	.669	.698	**.712**
pl_1	.681	.858	.870	**.875****	.707	.844	.848	**.855*****
pl_2	.619	.803	.814	**.825****	.697	.820	.834	**.839****
Avg.	.688	.828	.832	**.844**	.757	.806	.831	**.839**

(a) Results of the monolingual experiment for the WEI baseline, two reference methods (LR and KNN) as well as our FFNN model in Pearson r. Best result per data set and emotion format in bold, second best result underlined; significant difference (paired two-tailed t-test) over the second best system marked with “*”, “**”, or “***” for $p < .05$, .01, or .001, respectively.

	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dsg
en_1	.969	.741	.848	.962	.876	.871	.873	.805
en_2	.964	.704	.861	.942	.868	.821	.860	.799
es_1	.974	.771	.863	.957	.854	.833	.869	.752
es_2	.986	.828	.720	.977	.913	.867	.878	.807
es_3	.915	.692	—	.846	.839	.857	.842	.744
de_1	.929	.745	—	.894	.778	.644	.785	.461
pl_1	.963	.787	—	.946	.872	.826	.805	.826
pl_2	.947	.768	.760	.935	.844	.805	.790	.819
Avg.	.956	.754	.810	.932	.855	.816	.838	.752

(b) Results of the monolingual experiment per affective dimension in Pearson r. Color indicates outperforming human SHR (blue), being outperformed (red) or SHR not being reported (white; “—” meaning that the respective variable is not included).

Table 4: Results of the monolingual experiment.

As can be seen, all of the ERM approaches (LR, KNN, FFNN) perform more than 10%-points better than the state of the art in word emotion induction (WEI) for VAD prediction and at least about 5%-points better for BE5 predictions (on average over all data sets and affective variables). This finding already strongly suggests that ERM is the superior approach for automatic lexicon creation, given that the required data are available. This might be especially useful in situations where, say, large VAD but only small BE5 lexicons are available for a given language (see Section 4.4). Regarding the ordering of the ERM approaches, KNN outperforms LR in almost all cases. The advantage is more pronounced for mapping dim2cat (2.5%-points difference on average) than cat2dim (.4%-points difference). On top of that, our proposed FFNN model outperforms KNN by a 1.2%-point margin for cat2dim and a .8%-point margin for dim2cat (again as average over all data sets) performing best on each single data set. Regarding the 16 cases of Table 4a (8 data sets times two mapping directions), the performance gain of FFNN compared to the respective second best system is statistically significant[13] in all but 2 cases. The differences between the individual ERM approaches might appear quite small, yet become a lot more meaningful considering the proximity to human annotation capabilities as discussed in the following paragraphs.

Table 4b displays the performance figures of the FFNN model relative to each affective variable. As can be seen, among VAD, Valence is the easiest dimension to predict ($r = .956$ on average over all data sets) whereas for Arousal the performance is worst . Similarly, for BE5, Joy obtains the best values ($r = .932$) and Disgust is the hardest to predict. Interestingly, the overall ordering of performance within the two formats is consistent with the ordering of human reliability (see Table 3).

Comparing our system performance against human SHR (based on 20 participants per study; see Section 3.4), again our approach seems to be highly reliable (color coding of Table 4b). In particular, ERM using the FFNN model outperforms SHR in over half of the applicable cases (25 of 38). For mapping cat2dim it surpasses human reliability in all but 2 cases whereas when mapping dim2cat the reported SHR is surpassed in over half of the cases (14 out of 25).

This result, astonishing as it might appear, is yet consistent with findings from previous work which, in turn, were based on ISR (not on SHR) data (Buechel and Hahn, 2017a; Buechel and Hahn, 2018a). We conclude that in the monolingual set-up, ERM using the FFNN model substantially outperforms current capacities in word emotion induction and is even more reliable than a medium sized human rating study. Thus these automatically produced ratings should be cautiously attributed gold standard quality.

[13] Paired two-tailed t-tests based on the 10 train/test splits during cross-validation; $p < .05$.

4.3 Crosslingual Representation Mapping

In the crosslingual set-up, we make use of the fact that our model does not rely on any language-specific information, since the categories/dimensions describe supposedly universal affective states rather than linguistic entities. Thus, models trained on one language could, in theory, be applied to another one without any need for adaptation. This capability comes in handy when only data sets according to *one* emotion format exist for a given language. In such cases we could still train our model on data available for other languages and use it to produce new ratings for the language in focus. This section aims at estimating the performance of lexicons derived in this manner.

	Val	Aro	Joy	Ang	Sad	Fea	Dsg
en_1	.966	.683	.955	.858	.838	.817	.781
en_2	.956	.642	.934	.855	.810	.791	.800
es_1	.973	.692	.951	.786	.802	.782	.682
es_2	.985	.735	.974	.881	.860	.835	.787
es_3	.908	.548	.839	.821	.850	.807	.728
de_1	.927	.708	.889	.767	.618	.760	.458
pl_1	.957	.666	.937	.848	.784	.745	.801
pl_2	.938	.720	.932	.816	.785	.751	.809
Avg.	.951	.674	.926	.829	.793	.786	.731

Table 5: Results of crosslingual experiment in Pearson r. Color indicates outperforming human SHR (blue), being outperformed (red) or SHR not being reported (white).

For each of the data sets in Table 1, we trained FFNN models to map cat2dim and dim2cat, respectively. We trained on each gold lexicon that did not cover the language of the data set under scrutiny (e.g., for testing on en_1, the models were trained on all Spanish, Polish and German data sets, but not on en_2). Since this set-up leads to fixed train and test sets, we did not perform cross-validation. For comparability between data sets, the Dominance dimension was excluded for this experiment.

Overall, the results remained astonishingly stable compared to the monolingual set-up, with performance figures for Valence and Joy dropping by less than 1%-point on average over all data sets (see Table 5). Also, Anger, Sadness, Fear and Disgust only suffer a moderate decrease of about 5%-points at most—only the performance of Arousal decreased more than that.

A possible explanation for these strong results is the marked increase in the amount of training data that comes along with training on the majority of the available data (independent of language). This circumstance seems to counterbalance much of the negative effects that may arise in this crosslingual applications.

In comparison to SHR, the ERM approach still turns out to work quite well. Regarding VA, we outperform human reliability in 8 of 10 cases. Concerning BE5, SHR was beaten in about half of the cases (11 of 25). We conclude that, although the capability of our mapping approach suffers a bit in the crosslingual set-up, it still produces very accurate predictions and can thus be attested *near* gold quality, at least.

4.4 Automatic Lexicon Construction for Diverse Languages

After the positive evaluation of the FFNN model for ERM, the last bit of our contributions is to apply the created models to a wide variety of data sets which so far bear emotion ratings for *one* format only (either VA(D) or BE5). Based on the experiments reported so far, we claim that these have gold quality (for the monolingual approach, Section 4.2) or near-gold quality (for the crosslingual approach, Section 4.3).

For the monolingual approach, we train our model on the data set on which we achieved the highest performance in Section 4.2 for the respective language (assuming this hints at particularly “clean” data). In contrast, in the crosslingual set-up, training data are acquired by concatenating *all* the available data sets from Table 1 (consequently ignoring Dominance for compatibility).

Table 6 lists the emotion lexicons constructed in this manner together with their most important characteristics. The number of new ratings ranges from almost 13,000 (for English) and 10,500 (for Spanish), over several thousands (for Dutch, Chinese and Polish,) and around 1,500–1,000 (for Indonesian, Italian, Portuguese, Greek, French and German) to 200–100 (for Finnish and Swedish). For illustration, Table 2 displays three entries of the English BE5 lexicon, the largest one we constructed.

5 Conclusion

In this paper, we addressed the relatively new task of *emotion representation mapping*. It aims at transforming emotion ratings for lexical units from one emotion representation format into another one, e.g., mapping from Valence-Arousal-Dominance representations to Basic Emotion ones. Based on a large-scale evaluation we gathered solid empirical evidence that the proposed neural network model consistently outperforms the previous state-of-the-art performance figures in both word emotion induction and emotion representation mapping. Hence, the approach we propose currently constitutes the best-performing method for automatic emotion lexicon creation.

Mth	Lng	Format	Source	#Words
m	en	BE5	Warriner et al. (2013)	12,884
m	es	VAD	Stadthagen-González et al. (2017a)	10,489
m	de	BE5	Võ et al. (2009)	944
m	pl	BE5	Imbir (2016)	3,633
c	it	BE5	Montefinese et al. (2014)	1,121
c	pt	BE5	Soares et al. (2012)	1,034
c	nl	BE5	Moors et al. (2013)	4,299
c	id	BE5	Sianipar et al. (2016)	1,487
c	zh	BE5	Yu et al. (2016a); Yao et al. (2017)	3,797
c	fr	BE5	Monnier and Syssau (2014)	1,031
c	gr	BE5	Palogiannidi et al. (2016)	1,034
c	fn	BE5	Eilola and Havelka (2010)	210
c	sv	BE5	Davidson and Innes-Ker (2014)	99

Table 6: Overview of automatically constructed emotion lexicons; mapping methodology (monolingual or crosslingual), language (codes according to ISO 639-1), target emotion format, source lexicon of the mapping process and number of previously unknown ratings (excluding those present in other lexicons).

We also proposed a novel methodology for comparison against human rating capabilities based on normalized split-half reliability scores. For the first time, this allows for a large-scale evaluation against human performance. Our experimental data suggest that our models perform competitive relative to human assessments, even in cross-lingual applications, thus producing (near) gold quality data. We take this as a strong hint towards the reliability of the methods we propose.

Finally, we used these models to produce new emotion lexicons for 13 typologically diverse languages which are publicly available along with our code and experimental data (see Footnote 1).

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments and suggestions.

References

Jonathan Baxter. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Margaret M. Bradley and Peter J. Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA.

Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. Discrete Emotion Norms for Nouns: Berlin Affective Word List (DENN–BAWL). *Behavior Research Methods*, 43(2):441.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 1114–1122, The Hague, The Netherlands, August 29 – September 2, 2016.

Sven Buechel and Udo Hahn. 2017a. A flexible mapping scheme for discrete and dimensional emotion representations: Evidence from textual stimuli. In *CogSci 2017 — Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 180–185, London, UK, July 26–29, 2017.

Sven Buechel and Udo Hahn. 2017b. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, short papers, pages 578–585, Valencia, Spain, April 3–7, 2017.

Sven Buechel and Udo Hahn. 2018a. Representation mapping: A novel approach to generate high-quality multilingual emotion lexicons. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 184–191, Miyazaki, Japan, May 7–12, 2018.

Sven Buechel and Udo Hahn. 2018b. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, long papers, pages 1907–1918, New Orleans, Louisiana, USA, June 1–6, 2018.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Jean C. Carletta. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Per Davidson and Åse Innes-Ker. 2014. Valence and arousal norms for Swedish affective words. Technical Report Volume 14, No. 2, Lund University, Lund, Sweden.

Harris Drucker. 1997. Improving regressors using boosting techniques. In *ICML 1997 — Proceedings of the 14th International Conference on Machine Learning*, pages 107–115, Nashville, Tennessee, USA, July 8–12, 1997.

Steven Du and Xi Zhang. 2016. Aicyber's system for IALP 2016 Shared Task: Character-enhanced word vectors and boosted neural networks. In *IALP 2016 — Proceedings of the 2016 International Conference on Asian Language Processing*, pages 161–163, Tainan, Taiwan, November 21–23, 2016.

Tiina M. Eilola and Jelena Havelka. 2010. Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods*, 42(1):134–140.

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

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *CIKM 2005 — Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 617–624, Bremen, Germany, October 31 – November 05, 2005.

Pilar Ferré, Marc Guasch, Natalia Martínez-García, Isabel Fraga, and José Antonio Hinojosa. 2017. Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories. *Behavior Research Methods*, 49(3):1082–1094.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997 — Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July 7–12, 1997.

José Antonio Hinojosa, Natalia Martínez-García, Cristina Villalba-García, Uxia Fernández-Folgueiras, Alberto Sánchez-Carmona, Miguel Angel Pozo, and Pedro R. Montoro. 2016a. Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, 48(1):272–284.

José Antonio Hinojosa, Irene Rincón-Pérez, M. Verónica Romero-Ferreiro, Natalia Martínez-García, Cristina Villalba-García, Pedro R. Montoro, and Miguel Angel Pozo. 2016b. The Madrid Affective Database for Spanish (MADS): Ratings of dominance, familiarity, subjective age of acquisition and sensory experience. *PLoS One*, 11(5):e0155866.

Kurt Hornik. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257.

Kamil K. Imbir. 2016. Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, 7:#1081.

Diederik Kingma and Jimmy Ba. 2015. ADAM: A method for stochastic optimization. In *ICLR 2015 — Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15, San Diego, California, USA, May 7–9, 2015.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portorož, Slovenia, May 23–28, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS 2013 — Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada, USA, December 5–10, 2013.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In **SEM 2017 — Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 65–77, Vancouver, British Columbia, Canada, August 3–4, 2017.

Catherine Monnier and Arielle Syssau. 2014. Affective norms for French words (FAN). *Behavior Research Methods*, 46(4):1128–1137.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46(3):887–903.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbært. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45(1):169–177.

Elisavet Palogiannidi, Polychronis Koutsakis, Elias Iosif, and Alexandros Potamianos. 2016. Affective lexicon creation for the Greek language. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2867–2872, Portorož, Slovenia, 23–28 May 2016.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, October 25–29, 2014.

Ana P. Pinheiro, Marcelo Dias, João Pedrosa, and Ana P. Soares. 2017. Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*, 49(2):698–716.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL–R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

Sara Rosenthal, Preslav I. Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval 2015 Task 10: Sentiment analysis in Twitter. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*, pages 451–463, Denver, Colorado, USA, June 4–5, 2015.

Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu, and George Aaron Broadwell. 2016. ANEW+: Automatic expansion and validation of affective norms of words lexicons in multiple languages. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1127–1132, Portorož, Slovenia, 23–28 May 2016.

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, 7:#1907.

Ana Paula Soares, Montserrat Comesaña, Ana P. Pinheiro, Alberto Simões, and Carla Sofia Frade. 2012. The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1):256–269.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Hans Stadthagen-González, Pilar Ferré, Miguel A. Pérez-Sánchez, Constance Imbault, and José Antonio Hinojosa. 2017a. Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness. *Behavior Research Methods*. Available: `https://doi.org/10.3758/s13428-017-0962-y`.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez-Sánchez, and Marc Brysbært. 2017b. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1):111–123.

Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James. 2007. Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval 2007 Task 14: Affective text. In *SemEval 2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007*, pages 70–74, Prague, Czech Republic, June 23–24, 2007.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Melissa L. H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin Affective Word List Reloaded (BAWL–R). *Behavior Research Methods*, 41(2):534–538.

Henrica C. W. de Vet, Lidwine B. Mokkink, David G. Mosmuller, and Caroline B. Terwee. 2017. Spearman-Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85:45–49.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLoS One*, 10(7):e0132305.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4):1374–1385.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016a. Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California, USA, June 12–17, 2016.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016b. Overview of the IALP 2016 Shared Task on dimensional sentiment analysis for Chinese words. In *IALP 2016 — Proceedings of the 2016 International Conference on Asian Language Processing*, pages 156–160, Tainan, Taiwan, November 21–23, 2016.

12 Learning and Evaluating Emotion Lexicons for 91 Languages

Reference

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020a. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217.

Author Contributions

Udo Hahn performed supervision and project administration. Susanna Rücker performed software development and execution of experiments. I performed methodology and model development, data analysis and visualization, as well as supervised software development and execution of experiments. Experiments were designed jointly by Susanna Rücker and me. Conception and writing were performed jointly by all authors.

Learning and Evaluating Emotion Lexicons for 91 Languages

Sven Buechel, Susanna Rücker, and Udo Hahn
{sven.buechel|susanna.ruecker|udo.hahn}@uni-jena.de
Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
https://julielab.de

Abstract

Emotion lexicons describe the affective meaning of words and thus constitute a centerpiece for advanced sentiment and emotion analysis. Yet, manually curated lexicons are only available for a handful of languages, leaving most languages of the world without such a precious resource for downstream applications. Even worse, their coverage is often limited both in terms of the lexical units they contain and the emotional variables they feature. In order to break this bottleneck, we here introduce a methodology for creating almost arbitrarily large emotion lexicons for any target language. Our approach requires nothing but a source language emotion lexicon, a bilingual word translation model, and a target language embedding model. Fulfilling these requirements for 91 languages, we are able to generate representationally rich high-coverage lexicons comprising eight emotional variables with more than 100k lexical entries each. We evaluated the automatically generated lexicons against human judgment from 26 datasets, spanning 12 typologically diverse languages, and found that our approach produces results in line with state-of-the-art *monolingual* approaches to lexicon creation and even *surpasses human reliability* for some languages and variables. Code and data are available at github.com/JULIELab/MEmoLon archived under DOI 10.5281/zenodo.3779901.

1 Introduction

An emotion lexicon is a lexical repository which encodes the affective meaning of individual words (lexical entries). Most simply, affective meaning can be encoded in terms of *polarity*, i.e., the distinction whether an item is considered as positive, negative, or neutral. This is the case for many well-known resources such as WORDNET-AFFECT (Strapparava and Valitutti, 2004), SENTIWORDNET (Baccianella et al., 2010), or VADER (Hutto and Gilbert, 2014). Yet, an increasing number of researchers focus on more expressive encodings for affective states inspired by distinct lines of work in psychology (Yu et al., 2016; Buechel and Hahn, 2017; Sedoc et al., 2017; Abdul-Mageed and Ungar, 2017; Bostan and Klinger, 2018; Mohammad, 2018; Troiano et al., 2019).

Psychologists, on the one hand, value such lexicons as a controlled set of stimuli for designing experiments, e.g., to investigate patterns of lexical access or the structure of memory (Hofmann et al., 2009; Monnier and Syssau, 2008). NLP researchers, on the other hand, use them to augment the emotional loading of word embeddings (Yu et al., 2017; Khosla et al., 2018), as additional input to sentence-level emotion models so that the performance of even the most sophisticated neural network gets boosted (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; De Bruyne et al., 2019), or rely on them in a keyword-spotting approach when no training data is available, e.g., for studies dealing with historical language stages (Buechel et al., 2016).

As with any kind of manually curated resource, the availability of emotion lexicons is heavily restricted to only a few languages whose exact number varies depending on the variables under scrutiny. For example, we are aware of lexicons for 15 languages that encode the emotional variables of Valence, Arousal, and Dominance (see Section 2). This number leaves the majority of the world's (less-resourced) languages without such a dataset. In case such a lexicon exists for a particular language, it is often severely limited in size, sometimes only comprising some hundreds of entries (Davidson and Innes-Ker, 2014). Yet, even the largest lexicons typically cover only some ten thousands of words, still leaving out major portions of the emotion-carrying vocabulary. This is especially true for languages with complex morphology or

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1202–1217
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

productive compounding, such as Finnish, Turkish, Czech, or German. Finally, the diversity of emotion representation schemes adds another layer of complexity. While psychologists and NLP researchers alike find that different sets of emotional variables are complementary to each other (Stevenson et al., 2007; Pinheiro et al., 2017; Barnes et al., 2019; De Bruyne et al., 2019), *manually* creating emotion lexicons for every language and every emotion representation scheme is virtually impossible.

We here propose an approach based on cross-lingual distant supervision to generate almost arbitrarily large emotion lexicons for any target language and emotional variable, provided the following requirements are met: a source language emotion lexicon covering the desired variables, a bilingual word translation model, and a target language embedding model. By fulfilling these preconditions, we can *automatically* generate emotion lexicons for 91 languages covering ratings for eight emotional variables and hundreds of thousands of lexical entries each. Our experiments reveal that our method is on a par with state-of-the-art monolingual approaches and compares favorably with (sometimes even outperforms) human reliability.

2 Related Work

Representing Emotion. Whereas research in NLP has focused for a very long time almost exclusively on *polarity*, more recently, there has been a growing interest in more informative representation structures for affective states by including different groups of emotional variables (Bostan and Klinger, 2018). Borrowing from distinct schools of thought in psychology, these variables can typically be subdivided into *dimensional* vs. *discrete* approaches to emotion representation (Calvo and Mac Kim, 2013). The *dimensional* approach assumes that emotional states can be *composed* out of several foundational factors, most noticeably *Valence* (corresponding to polarity), *Arousal* (measuring calmness vs. excitement), and *Dominance* (the perceived degree of control in a social situation); VAD, for short (Bradley and Lang, 1994). Conversely, the *discrete* approach assumes that emotional states can be *reduced* to a small, evolutionary motivated set of basic emotions (Ekman, 1992). Although the exact division of the set has been subject of hot debates, recently constructed datasets (see Section 4) most often cover the categories of *Joy*, *Anger*, *Sadness*, *Fear*, and *Disgust*; BE5, for short. Plutchik's Wheel of Emotion takes a middle ground between those two positions by postulating emotional categories which are yet grouped into opposite pairs along different levels of intensity (Plutchik, 1980).

Another dividing line between representational approaches is whether target variables are encoded in terms of (strict) class-membership or scores for numerical strength. In the first case, emotion analysis translates into a (multi-class) classification problem, whereas the latter turns it into a regression problem (Buechel and Hahn, 2016). While our proposed methodology is agnostic towards the chosen emotion format, we will focus on the VAD and BE5 formats here, using numerical ratings (see the examples in Table 1) due to the widespread availability of such data. Accordingly, this paper treats word emotion prediction as a regression problem.

	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis
sunshine	8.1	5.3	5.4	4.2	1.2	1.3	1.3	1.2
terrorism	1.6	7.4	2.7	1.2	2.9	3.3	3.9	2.5
nuclear	4.3	7.3	4.1	1.4	2.2	1.9	3.2	1.6
ownership	5.9	4.4	7.5	2.1	1.4	1.2	1.4	1.3

Table 1: Sample entries from our English source lexicon described via eight emotional variables: **Val**ence, **Aro**usal, **Dom**inance [VAD], and **Joy**, **Ang**er, **Sad**ness, **Fea**r, and **Dis**gust [BE5]. VAD uses 1-to-9 scales ("5" encodes the neutral value) and BE5 1-to-5 scales ("1" encodes the neutral value).

Building Emotion Lexicons. Usually, the ground truth for affective word ratings (i.e., the assignment of emotional values to a lexical item) is acquired in a questionnaire study design where subjects (annotators) receive lists of words which they rate according to different emotion variables or categories. Aggregating individual ratings of multiple annotators then results in the final emotion lexicon (Bradley and Lang, 1999). Recently, this workflow has often been enhanced by crowdsourcing (Mohammad and Turney, 2013) and best-worst scaling (Kiritchenko and Mohammad, 2016).

As a viable alternative to manual acquisition, such lexicons can also be created by automatic means (Bestgen, 2008; Köper and Schulte im Walde, 2016; Shaikh et al., 2016), i.e., by learning to predict emotion labels for unseen words. Researchers have worked on this prediction problem for quite a long time. Early work tended to focus on word statistics, often in combination with linguistic rules (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003). More recent approaches focus heavily on word embeddings, either using semi-supervised graph-based approaches (Wang et al., 2016; Hamilton et al., 2016; Sedoc et al., 2017) or fully supervised methods (Rosenthal et al., 2015; Li et al., 2017; Rothe et al., 2016; Du and Zhang, 2016). Most important for this work, Buechel and Hahn (2018b) report on near-human performance using a combination of FASTTEXT vectors and a multi-task feed-forward network (see Section 4). While this line of work can add new words, it does not extend lexicons to other emotional variables or languages.

A relatively new way of generating novel labels is *emotion representation mapping* (ERM), an annotation projection that translates ratings from one emotion format into another, e.g., mapping VAD labels into BE5, or vice versa (Hoffmann et al., 2012; Buechel and Hahn, 2016, 2018a; Alarcão and Fonseca, 2017; Landowska, 2018; Zhou et al., 2020; Park et al., 2019). While our work uses ERM to add additional emotion variables to the source lexicon, ERM alone can neither increase the coverage of a lexicon, nor adapt it to another language.

Translating Emotions. The approach we propose is strongly tied to the observation by Leveau et al. (2012) and Warriner et al. (2013) who found—comparing a large number of existing emotion lexicons of different languages—that translational equivalents of words show strong stability and adherence to their emotional value. Yet, their work is purely descriptive. They do not exploit their observation to create new ratings, and only consider manual rather than automatic translation.

Making indirect use of this observation, Mohammad and Turney (2013) offer machine-translated versions of their *NRC Emotion Lexicon*. Also, many approaches in cross-lingual sentiment analysis (on the sentence-level) rely on translating polarity lexicons (Abdalla and Hirst, 2017; Barnes et al., 2018). Perhaps most similar to our work, Chen and Skiena (2014) create (polarity-only) lexicons for 136 languages by building a multilingual word graph and propagating sentiment labels through that graph. Yet, their method is restricted to high frequency words—their lexicons cover between 12 and 4,653 entries, whereas our approach exceeds this limit by more than two orders of magnitude.

Our methodology also resembles previous work which models word emotion for historical language stages (Cook and Stevenson, 2010; Hamilton et al., 2016; Hellrich et al., 2018; Li et al., 2019). Work in this direction typically comes up with a set of seed words with assumingly *temporally stable* affective meaning (our work assumes stability against translation) and then uses distributional methods to derive emotion ratings in the target language stage. However, gold data for the target language (stage) is usually inaccessible, often preventing evaluation against human judgment. In contrast, we here propose several alternative evaluation set-ups as an integral part of our methodology.

3 A Novel Approach to Lexicon Creation

Our methodology integrates (1) cross-lingual generation and expansion of emotion lexicons and (2) their evaluation against gold and silver standard data. Consequently, a key aspect of our workflow design is how data is split into train, dev, and test sets at different points of the generation process. Figure 1 gives an overview of our framework including a toy example for illustration.

Lexicon Generation. We start with a lexicon (`Source`) of arbitrary size, emotion format[1] and source language which is partitioned into train, dev, and test splits denoted by `Source-train`, `Source-dev`, and `Source-test`, respectively. Next, we leverage a bilingual word translation model between source and desired target language to build the first target-side emotion lexicon denoted as `TargetMT`. Source words are translated according to the model, whereas target-side emotion labels are simply copied from the source to the target (see Section 2). Entries are assigned to train, dev, or test set according to their source-side assignment (cf. Figure 1). The choice of our translation service (see below) ensures that each source word receives exactly one translation.

`TargetMT` is then used as the distant supervisor to train a model that predicts word emotions based on target-side word embeddings. `TargetMT-train` and `TargetMT-dev` are used to fit model parameters and optimize hyperparameters, respectively, whereas `TargetMT-test` is held out for later evaluation. Once finalized, the model is used to predict *new labels* for the words in `TargetMT`, resulting in a second target-side emotion lexicon denoted `TargetPred`. Our rationale for doing so is that a reasonably trained model should generalize well

[1] This encompasses not only VA(D) and BE5, but also any sort of (real-valued) polarity encodings.

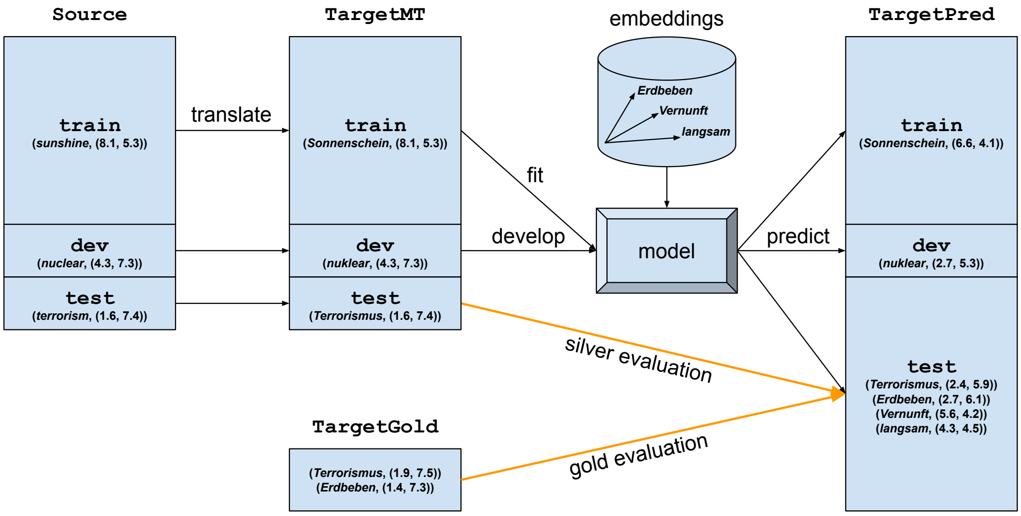


Figure 1: Schematic view on the methodology for generating and evaluating an emotion lexicon for a given target language based on source language supervision. Included is a toy example starting with an English VA lexicon (*sunshine, nuclear, terrorism* and the associated numerical scores for Valence and Arousal) and resulting in an extended German lexicon which incorporates translated entries with altered VA scores and additional entries originating from the embedding model with newly learned scores.

over the entire `TargetMT` lexicon because it has access to the target-side embedding vectors. Hence, it may mitigate some of the errors which were introduced in previous steps, either by machine translation or by assuming that source- and target-side emotion are always identical. We validate this assumption in Section 6. We also predict ratings for *all* the words in the embedding model, leading to a large number of new entries.

The splits are defined as follows: let MT_{train}, MT_{dev}, and MT_{test} denote the set of words in train, dev, and test split of `TargetMT`, respectively. Likewise, let P_{train}, P_{dev}, and P_{test} denote the splits of `TargetPred` and let E denote the set of words in the embedding model. Then

$$\begin{aligned} P_{\text{train}} &:= MT_{\text{train}} \\ P_{\text{dev}} &:= MT_{\text{dev}} \setminus MT_{\text{train}} \\ P_{\text{test}} &:= (MT_{\text{test}} \cup E) \setminus (MT_{\text{dev}} \cup MT_{\text{train}}) \end{aligned}$$

The above definitions help clarify the way we address polysemy.[2] Ambiguity on the target-side may result in multiple source entries translating to the same target-side word.[3] This circumstance leads to “partial duplicates” in `TargetMT`, i.e., groups of entries with the same word type but different emotion values (because they were derived from distinct `Source` entries). Such overlap could do harm to the integrity of our evaluation since knowledge may “leak” from training to validation phase, i.e., by testing the model on words it has already seen during training, although with distinct emotion labels. The proposed data partitioning eliminates such distortion effects. Since partial duplicates receive the same embedding vector, the prediction model assigns the same emotion value to both, thus merging them in `TargetPred`.

Evaluation Methodology. The main advantage of the above generation method is that it allows us to create large-scale emotion lexicons for languages

[2] In short, our work evades this problem by dealing with lexical entries exclusively on the type- rather than the sense-level. From a lexicological perspective, this may seem like a strong assumption. From a modeling perspective, however, it appears almost obvious as it aligns well with the major components of our methodology, i.e., lexicons, embeddings, and translation. The lexicons we work with follow the design of behavioral experiments: a stimulus (word type) is given to a subject and the response (rating) is recorded. The absence of sense-level annotation simplifies the mapping between lexicon and embedding entries. While sense embeddings form an active area of research (Camacho-Collados and Pilehvar, 2018; Chi and Chen, 2018), to the best of our knowledge, type-level embeddings yield state-of-the-art performance in downstream applications.

[3] Source-side polysemy, in contrast to its target-side counterpart, is less of a problem, because we receive only a single candidate during translation. This may result in cases where the translation misaligns with the copied emotion value in `TargetMT`. Yet, the prediction step partly mitigates such inconsistencies (see Section 6).

for which gold data is lacking. But if that is the case, how can we assess the quality of the generated lexicons? Our solution is to propose two different evaluation scenarios—a *gold evaluation* which is a strict comparison against human judgment, meaning that it is limited to languages where such data (denoted TargetGold) is available, and a *silver evaluation* which substitutes human judgments by automatically derived ones (silver standard) which is feasible for any language in our study. The rationale is that if both, gold and silver evaluation, strongly agree with each other, we can use one as proxy for the other when no target-side gold data exists (examined in Section 6).

Note that our lexicon generation approach consists of two major steps, *translation* and *prediction*. However, these two steps are not equally important for each generated entry in TargetPred. Words, such as German *Sonnenschein* for which a translational equivalent already exists in the Source ("sunshine"; see Figure 1), mainly rely on translation, while the prediction step acts as an optional refinement procedure. In contrast, the prediction step is crucial for words, such as *Erdbeben*, whose translational equivalents ("earthquake") are missing in the Source. Yet, these words also depend on the translation step for producing training data.

These considerations are important for deciding which words to evaluate on. We may choose to base our evaluation on the full TargetPred lexicon, including words from the training set—after all, the word emotion model does not have access to *any* target-side gold data. The problem with this approach is that it merges words that mainly rely on *translation*, because their equivalents are in the Source, and those which largely depend on *prediction*, because they are taken from the embedding model. In this case, generalizability of evaluation results becomes questionable.

Thus, our evaluation methodology needs to fulfill the following two requirements: (1) evaluation must not be performed on translational equivalents of the Source entries to which the model already had access during training (e.g., *Sonnenschein* and *nuklear* in our example from Figure 1); but, on the other hand, (2) a reasonable number of instances must be available for evaluation (ideally, as many as possible to increase reliability). The intricate cross-lingual train-dev-test set assignment of our generation methodology is in place so that we meet these two requirements.

ID	Encoding	Size	Citation
en1	VAD	1032	Warriner et al. (2013)
en2	VAD	1034	Bradley and Lang (1999)
en3	BE5	1034	Stevenson et al. (2007)
es1	VAD	1034	Redondo et al. (2007)
es2	VA	14031	Stadthagen-González et al. (2017)
es3	VA	875	Hinojosa et al. (2016)
es4	BE5	875	Hinojosa et al. (2016)
es5	BE5	10491	Stadthagen-González et al. (2018)
es6	BE5	2266	Ferré et al. (2017)
de1	VAD	1003	Schmidtke et al. (2014)
de2	VA	2902	Võ et al. (2009)
de3	VA	1000	Kanske and Kotz (2010)
de4	BE5	1958	Briesemeister et al. (2011)
pl1	VAD	4905	Imbir (2016)
pl2	VA	2902	Riegel et al. (2015)
pl3	BE5	2902	Wierzba et al. (2015)
zh1	VA	2794	Yu et al. (2016)
zh2	VA	1100	Yao et al. (2017)
it	VAD	1121	Montefinese et al. (2014)
pt	VAD	1034	Soares et al. (2012)
nl	VA	4299	Moors et al. (2013)
id	VAD	1487	Sianipar et al. (2016)
el	VAD	1034	Palogiannidi et al. (2016)
tr1	VA	2029	Kapucu et al. (2018)
tr2	BE5	2029	Kapucu et al. (2018)
hr	VA	3022	Ćoso et al. (2019)

Table 2: Lexicons used for gold evaluation. **ID**s consist of the respective ISO 639-1 language code plus a cardinal number to distinguish different datasets, if needed; the format of emotion **Encoding** is specified and **Size** gives the number of lexical entries per lexicon.

In particular, for our silver evaluation, we intersect TargetMT-test with TargetPred-test and compute the correlation of these two sets individually for each emotion variable. Pearson's r will be used as correlation measure throughout this paper. Establishing a test set at the very start of our workflow, Source-test, assures that there is a relatively large overlap between the two sets and, by extension, that our requirements for the evaluation are met.

The gold evaluation is a somewhat more challenging case, because we can, in general, not guarantee that the overlap of a TargetGold lexicon with TargetPred-test will be of any particular size. For this reason, the words of the embedding model are added to TargetPred-test (see above), maximizing the expected overlap with TargetGold. In practical terms, we intersect TargetGold with TargetPred-test and compute the variable-wise correlation between these sets, in parallel to the silver evaluation. A complementary strategy for maximizing overlap, by exploiting dependencies between published lexicons, is described below.

4 Experimental Setup

Gold Lexicons and Data Splits. We use the English emotion lexicon from Warriner et al. (2013) as first part of our `Source` dataset. This popular resource comprises about 14k entries in VAD format collected via crowdsourcing. Since manually gathered BE5 ratings are available only for a subset of this lexicon (Stevenson et al., 2007), we add BE5 ratings from Buechel and Hahn (2018a) who used emotion representation mapping (see Section 2) to convert the existing VAD ratings, showing that this is about as reliable as human annotation.

As apparent from the previous section, a crucial aspect for applying our methodology is the design of the train-dev-test split of the `Source` because it directly impacts the amount of words we can test our lexicons on during gold evaluation. In line with these considerations, we choose the lexical items which are already present in ANEW (Bradley and Lang, 1999) as `Source-test` set. ANEW is the precursor to the version later distributed by Warriner et al. (2013); it is widely used and has been adapted to a wide range of languages. With this choice, it is likely that a resulting `TargetPred-test` set has a large overlap with the respective `TargetGold` lexicon. As for the `TargetGold` lexicons, we included every VA(D) and BE5 lexicon we could get hold of with more than 500 entries. This resulted in 26 datasets covering 12 quite diverse languages (see Table 2). Note that we also include English lexicons in the gold evaluation. In these cases, no translation will be carried out (`Source` is identical to `TargetMT`) so that only the expansion step is validated. Appendix A.1 gives further details on data preparation.

Translation. We used the GOOGLE CLOUD TRANSLATION API[4] to produce word-to-word translation tables. This is a commercial service, total translation costs amount to 160 EUR. API calls were performed in November 2019.

Embeddings. We use the `fastText` embedding models from Grave et al. (2018) trained for 157 languages on the respective WIKIPEDIA and the respective part of COMMONCRAWL. These resources not only greatly facilitate our work but also increase comparability across languages. The restriction to "only" 91 languages comes from intersecting the ones covered by the vectors with the languages covered by the translation service.

Models. Since our proposed methodology is agnostic towards the chosen word emotion model, we will re-use models from the literature. In particular, we will rely on the multi-task learning feed-forward network (MTLFFN) worked out by Buechel and Hahn (2018b). This network constitutes the current state of the art for *monolingual* emotion lexicon creation (expanding an existing lexicon for a given language) for many of the datasets in Table 2.

The MTLFFN has two hidden layers of 256 and 128 units, respectively, and takes pre-trained embedding vectors as input. Its distinguishing feature is that hidden layer parameters are shared between the different emotion target variables, thus constituting a mild form of multi-task learning (MTL). We apply MTL to VAD and BE5 variables individually (but not between both groups), thus training two *distinct* emotion models per language, following the outcome of a development experiment. Details are given in Appendix A.2 together with the remainder of the model specifications.

Being aware of the infamous instability of neural approaches (Reimers and Gurevych, 2017), we also employ a ridge regression model, an L_2 regularized version of linear regression, as a more robust, yet also powerful baseline (Li et al., 2017).

5 Results

The size of the resulting lexicons (a complete list is provided in Table 8 in the Appendix) ranges from roughly 100k to more than 2M entries mainly depending on the vocabulary of the respective embeddings. We want to point out that not every single entry should be considered meaningful because of noise in the embedding vocabulary caused by typos and tokenization errors. However, choosing the "best" size for an emotion lexicon necessarily translates into a quality-coverage trade-off for which there is no general solution. Instead, we release the full-size lexicons and leave it to prospective users to apply any sort of filtering they deem appropriate.

Silver Evaluation. Figure 2 displays the results of our silver evaluation. Languages (x-axis) are sorted by their average performance over all variables (not shown in the plot; tabular data given in the Appendix). As can be seen, the evaluation results for English are markedly better than for any other language. This is not surprising since no (potentially error-prone) machine translation was performed. Apart from that, performance remains relatively stable across most of the languages and

[4] https://cloud.google.com/translate/

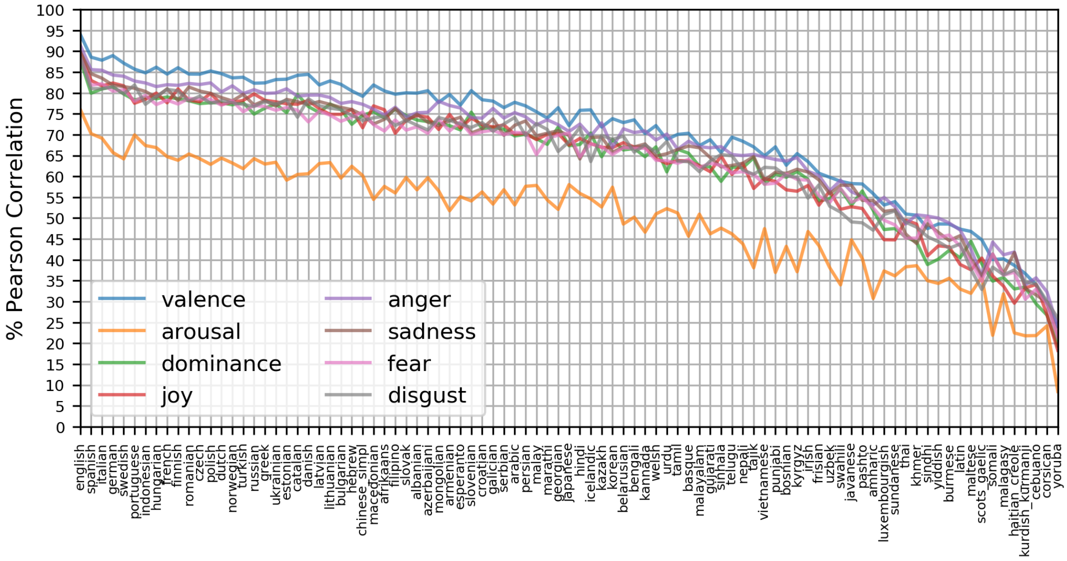


Figure 2: Silver evaluation results in Pearson's r. Languages (x-axis) are sorted according to mean correlation.

starts degrading more quickly only for the last third of them. In particular, for Valence—typically the easiest variable to predict—we achieve a strong performance of $r > .7$ for 56 languages. On the other hand, for Arousal—typically, the most difficult one to predict—we achieve a solid performance of $r > .5$ for 55 languages. Dominance and the discrete emotion variables show performance trajectories swinging between these two extremes. We assume that the main factors for explaining performance differences between languages are the quality of the translation and embedding models which, in turn, both depend on the amount of available text data (parallel or monolingual, respectively).

Comparing MTLFFN and ridge baseline, we find that the neural network reliably outperforms the linear model. On average over all languages and variables, the MTL models achieve 6.7%-points higher Pearson correlation. Conversely, ridge regression outperforms MTLFFN in only 15 of the total 728 cases (91 languages × 8 variables).

Gold Evaluation. Results for VAD variables on gold data are given in Table 3. As can be seen, our lexicons show a good correlation with human judgment and do so robustly, even for less-resourced languages, such as Indonesian (id), Turkish (tr), or Croatian (hr), and across affective variables. Perhaps the strongest negative outliers are the Arousal results for the two Chinese datasets (zh), which are likely to result from the low reliability of the gold ratings (see below).

ID	Shared	(%)	Val	Aro	Dom
en1	1032	100	**.94** (.87)	**.76** (.67)	**.88** (.76)
en2	1034	100	**.92** (.92)	.71 (**.73**)	.78 (**.82**)
es1	612	59	**.91** (.88)	**.71** (.70)	.82 (**.83**)
es2	7685	54	.79 (**.82**)	.64 (**.74**)	—
es3	363	41	.91	.73	—
de1	677	67	**.89** (.87)	.78 (**.80**)	.68 (**.74**)
de2	2329	80	.75	.64	—
de3	916	91	.80	.67	—
pl1	2271	46	**.83** (.74)	**.74** (.70)	.60 (**.69**)
pl2	1381	47	.82	.61	—
zh1	1685	60	.84 (**.85**)	.56 (**.63**)	—
zh2	701	63	.84	.44	—
it	660	58	**.89** (.86)	.63 (**.65**)	**.76** (.75)
pt	645	62	**.89** (.86)	.71 (**.71**)	**.75** (.73)
nl	2064	48	**.85** (.79)	.58 (**.74**)	—
id	696	46	**.84** (.80)	**.64** (.60)	**.63** (.58)
el	633	61	.86	.50	.74
tr1	721	35	.75	.57	—
hr	1331	44	.81	.66	—
Mn (all)			**.85**	**.65**	**.74**
Mn (vs. monolingual)			**.87** (.84)	.68 (**.70**)	**.74** (.74)

Table 3: Gold evaluation results for VAD (**Val**ence, **Aro**usal, **Dom**inance) in Pearson's r. Parentheses give comparative monolingual results from Buechel and Hahn (2018b). **Shared** words between `TargetGold` and `TargetPred-test`; **(%)**: percentage relative to `TargetGold`; **Mn** (all): mean over all datasets; **Mn** (vs. monolingual): mean over datasets with comparative results.

We compare these results against those from Buechel and Hahn (2018b) which were acquired on the respective `TargetGold` dataset in a monolingual fashion using 10-fold cross-validation (10-

ID	Shared	(%)	Joy	Ang	Sad	Fea	Dis
en3	1033	99	.89	.83	.80	.82	.78
es4	363	41	.86	.84	.84	.84	.76
es5	6096	58	.64	.72	.72	.72	.63
es6	992	43	.80	.74	.71	.72	.68
de4	848	43	.80	.66	.52	.68	.42
pl3	1381	47	.78	.71	.66	.69	.71
tr2	721	35	.77	.69	.71	.70	.65
Mean			**.79**	**.74**	**.71**	**.74**	**.66**

Table 4: Gold evaluation results for BE5 (**Joy**, **Ang**er, **Sad**ness, **Fea**r, **Dis**gust) in Pearson's r. **Shared** words between `TargetGold` and `TargetPred-test`; **(%)**: percentage relative to `TargetGold`; **Mean** over all datasets.

CV). We admit that those results are not fully comparable to those presented here because we use fixed splits rather than 10-CV. Nevertheless, we find that the results of our cross-lingual set-up are more than competitive, outperforming the monolingual results from Buechel and Hahn (2018b) in 17 out of 30 cases (mainly for Valence and Dominance, less often for Arousal). This is surprising since we use an otherwise identical model and training procedure. We conjecture that the large size of the English `Source` lexicon, compared to most `TargetGold` lexicons, more than compensates for error-prone machine translation.

Table 4 shows the results for BE5 datasets which are in line with the VAD results. Regarding the ordering of the emotional variables, again, we find Valence to be the easiest one to predict, Arousal the hardest, whereas basic emotions and Dominance take a middle ground.

Comparison against Human Reliability. We base this analysis on *inter-study reliability* (ISR), a rather strong criterion for human performance. ISR is computed, per variable, as the correlation between the ratings from two distinct annotation studies (Warriner et al., 2013). Hence, this analysis is restricted to languages where more than one gold lexicon exists per emotion format. We intersect the entries from both gold standards as well as the respective `TargetPred-test` set and compute the correlation between all three pairs of lexicons. If our lexicon agrees more with one of the gold standards than the two gold standards agree with each other, we consider this as an indicator for *superhuman* reliability (Buechel and Hahn, 2018b).

As shown in Table 5, our lexicons are often competitive with human reliability for Valence (especially for English and Chinese), but outperform human reliability in 4 out of 6 cases for Arousal, and in the single test case for Dominance. There are no cases of overlapping gold standards for BE5.

Gold1	Gold2	Shared	Emo	G1vsG2	G1vsPr	G2vsPr
en1	en2	1032	V	**.953**	.941	.922
			A	.760	**.761**	.711
			D	.794	**.879**	.782
es1	es2	610	V	**.976**	.905	.912
			A	**.758**	.714	.725
es2	es3	222	V	**.976**	.906	.907
			A	.710	**.724**	.691
de2	de3	498	V	**.963**	.806	.812
			A	**.760**	.721	.663
pl1	pl2	445	V	**.943**	.838	.852
			A	.725	**.764**	.643
zh1	zh2	140	V	**.932**	.918	.898
			A	.482	**.556**	.455

Table 5: Comparison against human performance. Correlation between two gold standards, **Gold1** and **Gold2**, with each other (**G1vsG2**), as well as with our lexicons `TargetPred-test` (**G1vsPr** and **G2vsPr**) relative to **Emo**tional variable and **Shared** number of words.

6 Methodological Assumptions Revisited

This section investigates patterns in prediction quality *across* languages, validating design decisions of our methodology.

Translation vs. Prediction. Is it beneficial to predict new ratings for the words in `TargetMT` rather than using them as final lexicon entries straight away? For each `TargetGold` lexicon (cf. Table 2), we intersect its word material with that in `TargetMT` and `TargetPred`. Then, we compute the correlation between `TargetPred` and `TargetMT` with the gold standard. This analysis was done on the respective *train* sets because using `TargetMT` rather than `TargetPred` is only an option for entries known at training time.

Table 6 depicts the results of this comparison averaged over all gold lexicons. As hypothesized, the `TargetPred` lexicons agree, on average, more with human judgment than the `TargetMT` lexicons, suggesting that the word emotion model acts as a value-adding post-processor, partly mitigating rating inconsistencies introduced by mere translation of the lexicons. The observation holds for each individual emotion variable with particularly large benefits for Arousal, where the post-processed `TargetPred` lexicons are on average

	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis
Pred	.871	.652	.733	.767	.734	.692	.728	.650
MT	.796	.515	.613	.699	.677	.636	.654	.579
Diff	.076	.137	.119	.068	.057	.056	.074	.071

Table 6: Quality of `TargetMT` vs. `TargetPred` in terms of average Pearson correlation over all languages and gold standards. Diff := `Pred` − `MT`.

14%-points better compared to the translation-only `TargetMT` lexicons. This seems to indicate that lexical Arousal is less consistent between translational equivalents compared to other emotional meaning components like Valence and Sadness, which appear to be more robust against translation.

Gold vs. Silver Evaluation. How meaningful is silver evaluation without gold data? We compute the Pearson correlation between gold and silver evaluation results across languages per emotion variable. For languages where we consider multiple datasets during gold evaluation, we first average the gold evaluation results for each emotion variable. As can be seen from Table 7, the correlation values range between $r = .91$ for Joy and $r = .27$ for Disgust. This relatively large dispersion is not surprising when we take into account that we correlate very small data series (for Valence and Arousal there are just 12 languages for which both gold and silver evaluation results are available; for BE5 there are only 5 such languages). However, the mean over all correlation values in Table 7 is .64, indicating that there is a relatively strong correlation between both types of evaluation. This suggests that the silver evaluation may be used as a rather reliable proxy of lexicon quality even in the absence of language-specific gold data.

	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis
#Lg	12	12	8	5	5	5	5	5
r	.54	.57	.52	.91	.85	.57	.87	.27

Table 7: Agreement between gold and silver evaluation across languages in Pearson's r relative to the number of applicable languages ("#Lg").

7 Conclusion

Emotion lexicons are at the core of sentiment analysis, a rapidly flourishing field of NLP. Yet, despite large community efforts, the coverage of existing lexicons is still limited in terms of languages, size, and types of emotion variables. While there are techniques to tackle these three forms of sparsity in isolation, we introduced a methodology which allows us to cope with them simultaneously by jointly combining emotion representation mapping, machine translation, and embedding-based lexicon expansion.

Our study is "large-scale" in many respects. We created representationally complex lexicons—comprising 8 distinct emotion variables—for 91 languages with up to 2 million entries each. The evaluation of the generated lexicons featured 26 manually annotated datasets spanning 12 diverse languages. The predicted ratings showed consistently high correlation with human judgment, compared favorably with state-of-the-art monolingual approaches to lexicon expansion and even surpassed human inter-study reliability in some cases.

The sheer number of test sets we used allowed us to validate fundamental methodological assumptions underlying our approach. Firstly, the evaluation procedure, which is integrated into the generation methodology, allows us to reliably estimate the quality of resulting lexicons, *even without target language gold standard*. Secondly, our data suggests that embedding-based word emotion models can be used as a *repair mechanism*, mitigating poor target-language emotion estimates acquired by simple word-to-word translation.

Future work will have to deepen the way we deal with word sense ambiguity by way of exchanging the simplifying type-level approach our current work is based on with a semantically more informed sense-level approach. A promising direction would be to combine a multilingual sense inventory such as BABELNET (Navigli and Ponzetto, 2012) with sense embeddings (Camacho-Collados and Pilehvar, 2018).

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions and comments, and Tinghui Duan, JULIE LAB, for assisting us with the Chinese gold data. This work was partially funded by the German Federal Ministry for Economic Affairs and Energy (funding line "Big Data in der makroökonomischen Analyse" [Big data in macroeconomic analysis]; Fachlos 2; GZ 23305/003#002).

References

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *IJCNLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 1: Long Papers, pages 506–515, Taipei, Taiwan, November 27 – December 1, 2017.

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 718–728, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

Soraia M. Alarcão and Manuel J. Fonseca. 2017. Identifying emotions in images from valence and arousal ratings. *Multimedia Tools and Applications*, 77(13):17413–17435.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2200–2204, La Valletta, Malta, May 17–23, 2010.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2483–2493, Melbourne, Victoria, Australia, July 15–20, 2018.

Jeremy Barnes, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2019. Lexicon information in neural sentiment analysis: a multi-task learning approach. In *NoDaLiDa 2019 — Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 175–186, Turku, Finland, September 30 – October 2, 2019.

Yves Bestgen. 2008. Building affective lexicons from specific corpora for automatic sentiment analysis. In *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 496–500, Marrakesh, Morocco, May 28–30, 2008.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August 20–26, 2018.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA.

Margaret M. Bradley and Peter J. Lang. 2010. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-2, University of Florida, Gainesville, Forida, USA.

Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. Discrete Emotion Norms for Nouns: Berlin Affective Word List (DENN–BAWL). *Behavior Research Methods*, 43(2):#441.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122, The Hague, The Netherlands, August 29 – September 2, 2016.

Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 578–585, Valencia, Spain, April 3–7, 2017.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904, Santa Fe, New Mexico, USA, August 20–26, 2018.

Sven Buechel and Udo Hahn. 2018b. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1: Long Papers, pages 1907–1918, New Orleans, Louisiana, USA, June 1–6, 2018.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the past: Adapting affective lexicons for historical emotion analysis. In *LT4DH 2016 — Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities @ COLING 2016*, pages 54–61, Osaka, Japan, December 11, 2016.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers, pages 383–389, Baltimore, Maryland, USA, June 23–25, 2014.

Ta-Chung Chi and Yun-Nung Chen. 2018. CLUSE : Cross-Lingual Unsupervised Sense Embeddings. In *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 271–281, Brussels, Belgium, October 31 – November 4, 2018.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 28–34, La Valletta, Malta, May 17–23, 2010.

Bojana Ćoso, Marc Guasch, Pilar Ferré, and José Antonio Hinojosa. 2019. Affective and concreteness norms for 3,022 Croatian words. *Quarterly Journal of Experimental Psychology*, 72(9):2302–2312.

Per Davidson and Åse Innes-Ker. 2014. Valence and arousal norms for Swedish affective words. *Lund Psychological Reports*, 14:#2.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2019. Joint emotion label space modelling for affect lexica. *arXiv:1911.08782 [cs]*.

Steven Du and Xi Zhang. 2016. Aicyber's system for IALP 2016 Shared Task: Character-enhanced word vectors and boosted neural networks. In *IALP 2016 — Proceedings of the 2016 International Conference on Asian Language Processing*, pages 161–163, Tainan, Taiwan, November 21–23, 2016.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Pilar Ferré, Marc Guasch, Natalia Martínez-García, Isabel Fraga, and José Antonio Hinojosa. 2017. Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories. *Behavior Research Methods*, 49(3):1082–1094.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3483–3487, Miyazaki, Japan, May 7–12, 2018.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas, USA, November 1–5, 2016.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997 — Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July 7–12, 1997.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JESEME: Interleaving semantics and emotions in a Web service for the exploration of language change phenomena. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, volume System Demonstrations, pages 10–14, Santa Fe, New Mexico, USA, August 20–26, 2018.

José Antonio Hinojosa, Natalia Martínez-García, Cristina Villalba-García, Uxia Fernández-Folgueiras, Alberto Sánchez-Carmona, Miguel Angel Pozo, and Pedro R. Montoro. 2016. Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, 48(1):272–284.

Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht-Ecklundt, Harald C. Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *SMC 2012 — Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3316–3320, Seoul, Korea, October 14–17, 2012.

Markus J. Hofmann, Lars Kuchinke, Sascha Tamm, Melissa L.-H. Võ, and Arthur M. Jacobs. 2009. Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience*, 9(4):389–397.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM 2014 — Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, pages 216–225, Ann Arbor, Michigan, USA, June 1–4, 2014.

Kamil K. Imbir. 2016. Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, 7:#1081.

Philipp Kanske and Sonja A. Kotz. 2010. Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, 42(4):987–991.

Aycan Kapucu, Aslı Kılıç, Yıldız Özkılıç, and Bengisu Sarıbaz. 2018. Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological Reports*, pages 1–22. [Available online Dec 4, 2018].

Sopan Khosla, Niyati Chhaya, and Kushal Chawla. 2018. AFF2VEC : Affect–enriched distributional word representations. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2204–2218, Santa Fe, New Mexico, USA, August 20–26, 2018.

Diederik Kingma and Jimmy Ba. 2015. ADAM: A method for stochastic optimization. In *ICLR 2015 — Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California, USA, May 7–9, 2015.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *NAACL-HLT 2016 — 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, USA, June 12–17, 2016.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portorož, Slovenia, May 23–28, 2016.

Agnieszka Landowska. 2018. Towards new mappings between emotion representation models. *Applied Sciences*, 8(2):#274.

Nicolas Leveau, Sandra Jhean-Larose, Guy Denhière, and Ba-Linh Nguyen. 2012. Validating an interlingual metanorm for emotional analysis of texts. *Behavior Research Methods*, 44(4):1007–1014.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Ying Li, Tomas Engelthaler, Cynthia S. Q. Siew, and Thomas T. Hills. 2019. The MACROSCOPE: A tool for examining the historical structure of language. *Behavior Research Methods*, 51(4):1864–1877.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 174–184, Melbourne, Victoria, Australia, July 15–20, 2018.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *WASSA 2017 — Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2017*, pages 34–49, Copenhagen, Denmark, September 8, 2017.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SEMEVAL-2018 Task 1: Affect in Tweets. In *SemEval 2018 — Proceedings of the 12th International Workshop on Semantic Evaluation @ NAACL-HLT 2018*, pages 1–17, New Orleans, Louisiana, USA, June 5–6, 2018.

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

Catherine Monnier and Arielle Syssau. 2008. Semantic contribution to verbal short-term memory: Are pleasant words easier to remember than neutral words in serial recall and serial recognition? *Memory & Cognition*, 36(1):35–42.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46(3):887–903.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbært. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45(1):169–177.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BABELNET: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Elisavet Palogiannidi, Polychronis Koutsakis, Elias Iosif, and Alexandros Potamianos. 2016. Affective lexicon creation for the Greek language. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2867–2872, Portorož, Slovenia, May 23–28, 2016.

Sungjoon Park, Jiseon Kim, Jaeyeol Jeon, Heeyoung Park, and Alice Oh. 2019. Toward dimensional emotion detection from categorical emotion annotations. *arXiv:1911.02499 [cs, eess]*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. SCIKIT-LEARN: Machine learning in PYTHON. *Journal of Machine Learning Research*, 12(85):2825–2830.

Ana P. Pinheiro, Marcelo Dias, João Pedrosa, and Ana P. Soares. 2017. Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*, 49(2):698–716.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Emotion: Theory, Research and Experience*, volume 1: Theories of Emotion, pages 3–33. Academic Press, New York, NY, USA.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September 9–11, 2017.

Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL–R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

Sara Rosenthal, Preslav I. Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SEMEVAL 2015 Task 10: Sentiment Analysis in Twitter. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*, pages 451–463, Denver, Colorado, USA, June 4–5, 2015.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California, USA, June 12–17, 2016.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. 2014. ANGST: Affective Norms for German Sentiment Terms, derived from the Affective Norms for English Words. *Behavior Research Methods*, 46(4):1108–1118.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle H. Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 564–571, Valencia, Spain, April 3–7, 2017.

Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu, and George Aaron Broadwell. 2016. ANEW+ : Automatic expansion and validation of Affective Norms of Words lexicons in multiple languages. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1127–1132, Portorož, Slovenia, May 23–28, 2016.

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, 7:#1907.

Ana Paula Soares, Montserrat Comesaña, Ana P. Pinheiro, Alberto Simões, and Carla Sofia Frade. 2012. The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1):256–269.

Hans Stadthagen-González, Pilar Ferré, Miguel A. Pérez-Sánchez, Constance Imbault, and José Antonio Hinojosa. 2018. Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness. *Behavior Research Methods*, 50(5):1943–1952.

Hans Stadthagen-González, Constance Imbault, Miguel A. Pérez-Sánchez, and Marc Brysbært. 2017. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1):111–123.

Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James. 2007. Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Carlo Strapparava and Alessandro Valitutti. 2004. WORDNET-AFFECT: An affective extension of WORDNET. In *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, Lisbon, Portugal, May 24–30, 2004.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy, July 28 – August 2, 2019.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Melissa L.-H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin Affective Word List Reloaded (BAWL–R). *Behavior Research Methods*, 41(2):534–538.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957–1968.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLoS ONE*, 10(7):#e0132305.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4):1374–1385.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California, USA, June 12–17, 2016.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark, September 9–11, 2017.

Feng Zhou, Shu Kong, Charless C. Fowlkes, Tao Chen, and Baiying Lei. 2020. Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*. [Available online Jan 23, 2020].

A Appendices

A.1 Data Preparation

The exact design of the Source train-dev-test split is as follows: All entries (words plus ratings) from all splits are taken from Warriner et al. (2013). The data was then partitioned based on the overlap with the two precursory versions by Bradley and Lang (1999) (the original ANEW) and Bradley and Lang (2010) (an early extended version of ANEW roughly twice as large). Source-test was built by intersecting the lexicon from Warriner et al. (2013) with the original ANEW. A similar process was applied for Source-dev: we intersected the words from Warriner et al. (2013) and Bradley and Lang (2010) and removed the ones present in Source-test. Lastly, Source-train is made up by all words from Warriner et al. (2013) which are neither in Source-test nor in Source-dev. The reason why the ratings in Source are taken exclusively from Warriner et al. (2013) is that these are distributed under a more permissive license compared to their precursors.

We removed multi-token entries (e.g., *boa constrictor*) and entries with upper case characters (e.g., *Budweiser*) from all data splits of Source, thus restricting the lexicon to single-token, non-proper noun entries to make it more suitable for word embedding-based research. All splits combined have 13,791 entries (train: 11,463, dev: 1,296, test: 1,032), thus removing less than 1% from the original lexicon.[5]

Regarding the remaining gold standards, the only cases which needed additional preparation or cleansing steps were zh1 (Yu et al., 2016) and zh2 (Yao et al., 2017). zh1 was created and is distributed using traditional Chinese characters, whereas the embedding model by Grave et al. (2018) employs simplified ones. Therefore, we converted zh1 into simplified characters using GOOGLE TRANSLATE[6] prior to evaluation.

While manually examining the zh2 lexicon, we noticed several cases where the ratings seemed rather counter-intuitive (e.g., seemingly positive words which received very negative ratings). We contacted the authors who confirmed the problem and sent us a corrected version. We did not find any such problems in the second version. We consulted with a Chinese native speaker for both of these procedures regarding the zh1 and zh2 lexicons.

A.2 Model Training and Implementation

Training of the MTLFFN model closely followed the procedure specified by Buechel and Hahn (2018b): For each language, the model was trained for roughly 15k iterations (exactly 168 epochs) with a batch size of 128 using the Adam optimizer (Kingma and Ba, 2015) with learning rate 10^{-3}, and .5 dropout on the hidden layers and .2 on the input layer. As nonlinear activation function we used leaky ReLU with “leakage” of 0.01.

Embedding vectors are the only model input. They have 300 dimensions for every language, independent of their respective training data size (Grave et al., 2018). Since the automatic translation of Source is not guaranteed to result in single-word translations, we use the following workaround to derive embedding vectors for multi-token translations: If the translation as a whole cannot be found in the embedding model, the multi-token term gets split up into its constituent parts, using spaces, apostrophes or hyphens as separators. Each substring is looked up in the embedding model, the averaged vector is taken as input. If no substring is recognized, we use the zero vector instead. We also use the zero vector for single-token entries in TargetMT that are missing in the embeddings.

Since Buechel and Hahn (2018b) considered only VAD but not BE5 datasets, we conducted a development experiment on the TargetMT-dev sets for all 91 languages where we assessed whether MTL is advantageous for BE5 variables as well, or for a combination of VAD and BE5 variables. We found that MTL improved performance when applied separately among all VAD and BE5 variables. Yet, when jointly learning all eight emotion variables, the results were somewhat inconclusive. Performance *increased* for BE5, but *decreased* for VAD. Hence, for lexicon creation, we took a cautious approach and trained *two separate models per language*, one for VAD, the other for BE5. An analysis of MTL across VAD and BE5 is left for future work.

The MTLFFN model is implemented in PYTORCH, adapting part of the TENSORFLOW code from Buechel and Hahn (2018b). The ridge regression baseline model is implemented with SCIKIT-LEARN (Pedregosa et al., 2011) using default parameters.

[5] The data split is available at: https://github.com/JULIELab/XANEW

[6] In this case the regular Web application, not the API, was used: https://translate.google.com/

No.	ISO	Full Name	Size	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis	Mean
1	en	English	2,000,004	.94	.76	.88	.90	.91	.90	.89	.89	.88
2	es	Spanish	2,001,183	.89	.70	.80	.83	.86	.85	.82	.81	.82
3	it	Italian	2,001,137	.88	.69	.81	.82	.85	.84	.82	.81	.81
4	de	German	2,000,507	.89	.66	.81	.82	.84	.82	.80	.81	.81
5	sv	Swedish	2,000,980	.87	.64	.80	.82	.84	.82	.81	.80	.80
6	pt	Portuguese	2,001,078	.86	.70	.78	.78	.83	.81	.78	.82	.79
7	id	Indonesian	2,002,221	.85	.67	.79	.78	.82	.80	.79	.77	.79
8	hu	Hungarian	2,000,975	.86	.67	.79	.80	.82	.79	.77	.79	.79
9	fr	French	2,001,517	.85	.65	.79	.78	.82	.81	.78	.81	.78
10	fi	Finnish	2,000,841	.86	.64	.79	.81	.82	.78	.77	.80	.78
11	ro	Romanian	2,001,501	.85	.65	.78	.78	.82	.81	.79	.78	.78
12	cs	Czech	2,001,203	.84	.64	.77	.78	.82	.80	.79	.79	.78
13	pl	Polish	2,001,460	.85	.63	.78	.80	.82	.80	.78	.78	.78
14	nl	Dutch	2,000,721	.85	.64	.78	.77	.80	.79	.77	.78	.77
15	no	Norwegian (Bokmål)	2,000,876	.84	.63	.77	.78	.82	.78	.78	.78	.77
16	tr	Turkish	2,002,489	.84	.62	.78	.78	.80	.80	.75	.77	.77
17	ru	Russian	2,001,317	.82	.64	.75	.80	.81	.77	.77	.77	.77
18	el	Greek	2,001,704	.82	.63	.76	.78	.80	.78	.77	.78	.77
19	uk	Ukrainian	2,001,261	.83	.63	.77	.78	.80	.77	.76	.77	.76
20	et	Estonian	2,001,125	.83	.59	.75	.77	.81	.78	.77	.78	.76
21	ca	Catalan	2,001,538	.84	.60	.80	.77	.79	.78	.76	.74	.76
22	da	Danish	2,000,654	.84	.61	.77	.78	.79	.77	.73	.79	.76
23	lv	Latvian	1,642,923	.82	.63	.75	.76	.79	.78	.76	.77	.76
24	lt	Lithuanian	2,001,306	.83	.63	.77	.75	.79	.77	.75	.76	.76
25	bg	Bulgarian	2,001,391	.82	.60	.76	.75	.77	.77	.73	.76	.74
26	he	Hebrew	2,001,984	.80	.62	.72	.76	.78	.76	.74	.75	.74
27	zh	Chinese	2,001,799	.79	.60	.75	.72	.77	.75	.75	.73	.73
28	mk	Macedonian	1,356,402	.82	.54	.75	.77	.76	.73	.72	.74	.73
29	af	Afrikaans	883,464	.80	.58	.74	.76	.75	.74	.71	.74	.73
30	tl	Tagalog	716,272	.80	.56	.76	.70	.77	.76	.74	.72	.73
31	sk	Slovak	2,001,221	.80	.60	.75	.74	.74	.73	.71	.73	.72
32	sq	Albanian	1,169,697	.80	.57	.73	.75	.75	.75	.72	.72	.72
33	az	Azerbaijani	2,002,146	.81	.60	.73	.74	.75	.73	.70	.71	.72
34	mn	Mongolian	608,598	.78	.57	.73	.71	.78	.72	.74	.74	.72
35	hy	Armenian	2,001,329	.80	.52	.72	.75	.77	.73	.71	.73	.72
36	eo	Esperanto	2,001,575	.77	.55	.71	.72	.76	.74	.73	.73	.71
37	sl	Slovenian	1,992,272	.81	.54	.75	.74	.74	.70	.70	.72	.71
38	hr	Croatian	2,001,570	.78	.56	.71	.72	.74	.71	.71	.73	.71
39	gl	Galician	1,336,256	.78	.53	.72	.72	.76	.74	.71	.71	.71
40	sr	Serbian	2,002,395	.76	.57	.71	.72	.74	.70	.70	.73	.70
41	ar	Arabic	2,003,155	.78	.53	.70	.70	.75	.72	.71	.74	.70
42	fa	Persian	2,003,533	.77	.58	.70	.70	.74	.73	.70	.70	.70
43	ms	Malay	1,213,397	.75	.58	.69	.69	.72	.70	.65	.73	.69
44	mr	Marathi	848,549	.74	.54	.68	.70	.74	.70	.69	.71	.69
45	ka	Georgian	1,567,232	.76	.52	.72	.70	.72	.71	.70	.66	.69
46	ja	Japanese	2,003,306	.72	.58	.67	.68	.71	.70	.70	.68	.68
47	hi	Hindi	1,879,196	.76	.56	.68	.69	.73	.64	.65	.72	.68
48	is	Icelandic	945,214	.76	.55	.70	.68	.70	.69	.68	.64	.67
49	kk	Kazakh	1,981,562	.72	.53	.65	.67	.73	.69	.67	.70	.67
50	ko	Korean	2,002,600	.74	.57	.69	.67	.67	.66	.65	.69	.67
51	be	Belarusian	1,715,582	.73	.49	.66	.68	.71	.67	.67	.70	.66
52	bn	Bengali	1,471,709	.74	.50	.67	.67	.70	.67	.67	.66	.66
53	kn	Kannada	1,747,421	.70	.47	.65	.67	.71	.68	.67	.68	.65
54	cy	Welsh	502,006	.72	.51	.67	.64	.69	.65	.64	.66	.65
55	ur	Urdu	1,157,969	.69	.52	.61	.63	.70	.65	.64	.68	.64
56	ta	Tamil	2,002,514	.70	.51	.66	.64	.66	.66	.63	.64	.64
57	eu	Basque	1,828,013	.70	.46	.66	.64	.68	.67	.64	.64	.64
58	ml	Malayalam	2,002,920	.67	.51	.62	.63	.67	.67	.62	.61	.63
59	gu	Gujarati	557,270	.69	.46	.62	.61	.67	.65	.63	.64	.62
60	si	Sinhalese	812,356	.66	.48	.59	.65	.67	.62	.63	.65	.62
61	te	Telugu	1,880,585	.69	.46	.62	.60	.65	.63	.61	.65	.61
62	ne	Nepali	580,582	.68	.44	.62	.63	.65	.63	.61	.62	.61
63	tg	Tajik	508,617	.67	.38	.64	.57	.65	.65	.60	.60	.60
64	vi	Vietnamese	2,008,605	.65	.47	.58	.59	.65	.59	.58	.62	.59
65	pa	Eastern Punjabi	403,997	.67	.37	.61	.59	.64	.61	.58	.62	.59
66	bs	Bosnian	1,124,938	.63	.43	.60	.57	.64	.61	.61	.60	.58
67	ky	Kirghiz	751,902	.65	.37	.61	.56	.64	.62	.59	.60	.58
68	ga	Irish	321,249	.64	.47	.59	.58	.61	.61	.59	.55	.58
69	fy	West Frisian	530,054	.61	.43	.54	.53	.60	.59	.55	.58	.56
70	uz	Uzbek	833,860	.60	.38	.55	.56	.57	.56	.54	.53	.53
71	sw	Swahili	391,312	.59	.34	.57	.52	.59	.58	.57	.51	.53
72	jv	Javanese	518,634	.58	.45	.53	.53	.56	.58	.54	.49	.53
73	ps	Pashto	300,927	.58	.40	.56	.52	.55	.54	.55	.49	.53
74	am	Amharic	308,109	.56	.31	.52	.48	.53	.54	.52	.47	.49
75	lb	Luxembourgish	642,504	.53	.37	.47	.45	.55	.52	.50	.51	.49
76	su	Sundanese	327,533	.54	.36	.47	.45	.53	.52	.48	.52	.48
77	th	Thai	2,006,540	.51	.38	.45	.50	.49	.46	.45	.49	.47
78	km	Khmer	247,498	.51	.39	.44	.49	.51	.44	.45	.48	.46
79	sd	Sindhi	139,063	.47	.35	.39	.41	.50	.49	.50	.46	.45
80	yi	Yiddish	205,727	.49	.34	.40	.43	.50	.47	.45	.44	.44
81	my	Burmese	339,628	.49	.36	.42	.43	.49	.45	.46	.43	.44
82	la	Latin	1,088,139	.47	.33	.40	.39	.47	.46	.43	.44	.42
83	mt	Maltese	204,630	.47	.32	.44	.38	.43	.40	.39	.38	.40
84	gd	Scottish Gaelic	150,694	.45	.36	.39	.40	.36	.36	.35	.33	.38
85	so	Somali	177,405	.40	.22	.35	.36	.44	.41	.41	.38	.37
86	mg	Malagasy	415,050	.40	.32	.36	.34	.41	.37	.36	.36	.37
87	ht	Haitian	118,302	.39	.22	.33	.30	.42	.42	.37	.38	.35
88	ku	Kurdish (Kurmanji)	395,645	.37	.22	.33	.33	.34	.33	.31	.35	.32
89	ceb	Cebuano	2,006,001	.34	.22	.29	.34	.36	.32	.33	.34	.32
90	co	Corsican	108,035	.29	.24	.27	.27	.32	.30	.29	.30	.29
91	yo	Yoruba	156,764	.24	.08	.19	.18	.24	.21	.21	.26	.20

Table 8: Overview of generated emotion lexicons with silver evaluation results; sorted by **Mean** performance over the eight emotional variables.

13 Towards Label-Agnostic Emotion Embeddings

Reference

Sven Buechel, Luise Modersohn, and Udo Hahn. 2021. Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249.

Author Contributions

Udo Hahn performed supervision and project administration. Luise Modersohn performed model and software development. I performed methodology, model, and software development, experimental design and execution of experiments, data analysis and visualization. Conception and writing were performed jointly by all authors.

Towards Label-Agnostic Emotion Embeddings

Sven Buechel **Luise Modersohn** **Udo Hahn**
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Fürstengraben 27, 07743 Jena, Germany
firstname.lastname@uni-jena.de
https://julielab.de

Abstract

Research in emotion analysis is scattered across different label formats (e.g., polarity types, basic emotion categories, and affective dimensions), linguistic levels (word vs. sentence vs. discourse), and, of course, (few well-resourced but much more under-resourced) natural languages and text genres (e.g., product reviews, tweets, news). The resulting heterogeneity makes data and software developed under these conflicting constraints hard to compare and challenging to integrate. To resolve this unsatisfactory state of affairs we here propose a training scheme that learns a shared latent representation of emotion independent from different label formats, natural languages, and even disparate model architectures. Experiments on a wide range of datasets indicate that this approach yields the desired interoperability without penalizing prediction quality. Code and data are archived under DOI 10.5281/zenodo.5466068.

1 Introduction

Emotion analysis in the field of NLP[1] has experienced a remarkable evolution of representation schemes. Starting from the early focus on *polarity*, i.e., the main distinction between positive and negative feelings emerging from natural language utterances (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003), the number and variety of label formats, i.e., groups of emotional target variables and their associated value ranges, has been growing rapidly (Bostan and Klinger, 2018; De Bruyne et al., 2020). This development is a double-edged sword though.

On the one hand, the wide variety of available label formats allows NLP models to become more informative and richer in expressive power. This gain is because many of the newer representation schemes follow well-researched branches of psychological theory, such as basic emotion categories or affective dimensions (Ekman, 1992; Russell and Mehrabian, 1977), which offer information complementary to each other (Stevenson et al., 2007). Others argue that different emotional nuances turn out to be particularly useful for specific targeted downstream applications (Bollen et al., 2011; Desmet and Hoste, 2013).

On the other hand, this proliferation of label formats has led to a severe loss in cross-data comparability. As Tab. 1 illustrates, the total volume of available gold data is spread not only over distinct languages but also a huge number of emotion annotation schemes. Consequently, comparing or even merging data from different rating studies is often impossible. This, in turn, contributes to the development of an unnecessarily large number of prediction models, each with limited coverage of the full range of human emotion.

To escape from these dilemmata, we propose a method that mediates between such different representation schemes. In contrast to previous work which unified *some* sources of heterogeneity (see §2), to the best of our knowledge, our approach is the first to learn a representation space for emotions that *generalizes* over individual languages, emotion label formats, and distinct model architectures for emotion analysis.

Technically speaking, our approach consists of a set of pre-trained prediction heads that can be easily attached to existing state-of-the-art neural models. Doing so, a model learns to *embed* language items of a particular domain in a shared representation space that resembles an “interlingua for emotion”. These “emotion embeddings” capture a rich array of affective nuances and allow for a direct comparison of emotional load between heterogeneous samples (see Fig. 1). They may thus form a solid basis for a broad range of linguistic, psychological, and cultural follow-up studies.

[1] We use “emotion” as an umbrella term for phenomena such as polarity, sentiment, feelings, or affective states.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9231–9249
November 7–11, 2021. ©2021 Association for Computational Linguistics

Sample	Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis
rollercoaster	8.0○	8.1○	5.1○	3.4□	1.4□	1.1□	2.8□	1.1□
urine	3.3○	4.2○	5.2○	1.9□	1.4 □	1.2□	1.4□	2.6□
szczęśliwy [(a)]	2.8●	4.0○						
College tution continues climbing				0■	54■	40■	3■	31■
A gentle, compassionate drama about grief and healing	*pos*△							
喇叭這一代還是差勁透了。[(b)]	2.8○	6.1○						

Value Ranges: ○[1, 9] ●[−3, 3] △{*pos*, *neg*} □[1, 5] ■[0, 100]

Table 1: Sample entries from various sources described along eight emotional variables:
[VAD]—**Val**ence (≈ **Pol**arity), **Aro**usal, **Dom**inance, and [BE5]—**Joy**, **Ang**er, **Sad**ness, **Fea**r, and **Dis**gust.
Samples differ in languages addressed (English, Polish, Mandarin), linguistic domain (word vs. text, register) and label format (covered variables and their value ranges).
Translations: [(a)] "happy" (from Polish); [(b)] "This product generation still has terrible speakers." (from Mandarin)

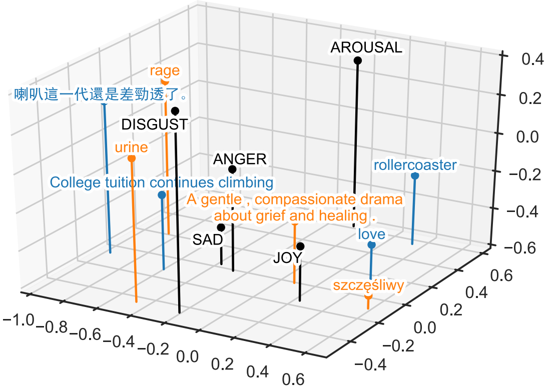


Figure 1: Emotional loading of heterogenous samples in common representation space with selected emotion variables (in capitals); first three principal components. Color only used as visual aid. Translations for non-English items are given in Tab. 1.

In terms of practical benefits, our method allows models to predict label formats unseen during training and lowers space requirements by reducing a large number of format-specific models to a small number of format-agnostic ones. Although not in the center of interest of this study, our approach also often leads to small improvements in prediction quality, as experiments on 13 datasets for 6 natural languages reveal.

2 Related Work

Representing Emotion. At the heart of computational emotion representation lies a set of *emotion variables* ("classes", "constructs") used to capture different facets of affective meaning. Researchers may choose from a multitude of approaches designed in the long and controversial history of the psychology of emotion (Scherer, 2000; Hofmann et al., 2020). A popular choice are so-called *basic emotions* (Alm et al., 2005; Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007), such as the six categories identified by Ekman (1992): *Joy*, *Anger*, *Sadness*, *Fear*, *Disgust*, and *Surprise* (**BE6**, for short). A subset of these excluding *Surprise* (**BE5**) is often used for emotional word datasets in psychology ("affective norms") which are available for a wide range of languages.

Affective dimensions constitute a popular alternative to basic emotions (Yu et al., 2016; Sedoc et al., 2017; Buechel and Hahn, 2017; Li et al., 2017; Mohammad, 2018). The most important ones are *Valence* (negative vs. positive, thus corresponding to the notion of *polarity*; Turney and Littman, 2003) and *Arousal* (calm vs. excited) (**VA**). These two dimensions are sometimes extended by *Dominance* (feeling powerless vs. empowered; **VAD**).

Other theories influential for NLP include Plutchik's (2001) *Wheel of Emotion* (Mohammad and Turney, 2013; Abdul-Mageed and Ungar, 2017; Tafreshi and Diab, 2018; Bostan et al., 2020) and appraisal dimensions (Balahur et al., 2012; Troiano et al., 2019; Hofmann et al., 2020). Yet frequently, studies do not follow any of these established approaches but rather design a customized set of variables in an ad-hoc fashion, often driven by the availability of user-labeled data in social media, or the specifics of an application or domain which requires attention to particular emotional nuances (Bollen et al., 2011; Desmet and Hoste, 2013; Staiano and Guerini, 2014; Qadir and Riloff, 2014; Li et al., 2016; Demszky et al., 2020).

This proliferating diversity of emotion label formats is the reason for the lack of comparability outlined in §1. Our work aims to unify these heterogeneous labels by learning to translate them into a shared distributional representation (see Fig. 1).

Analyzing Emotion. There are several subtasks in emotion analysis that require distinct model types. Word-level prediction (or "emotion lexicon induction") is concerned with the emotion associated with an individual word out of context. Early work exploited primarily surface patterns of word usage (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003) whereas more recent activities rely on more sophisticated statistical signals encoded in word embeddings (Amir et al., 2015; Rothe et al., 2016; Li et al., 2017). Combinations of high-quality embeddings with feed-forward nets have proven to be very successful, rivaling human annotation capabilities (Buechel and Hahn, 2018b).

In contrast, modeling emotion of sentences or short texts (jointly referred to as "text") was traditionally based largely on lexical resources (Taboada et al., 2011). Later, those were combined with conventional machine learning techniques (Mohammad et al., 2013) before being widely replaced by neural end-to-end approaches (Socher et al., 2013; Kim, 2014; Abdul-Mageed and Ungar, 2017). Current state-of-the-art results are achieved by transfer learning with transformer models (Devlin et al., 2019; Zhong et al., 2019; Delbrouck et al., 2020).

Our work complements these lines of research by providing a method that allows existing models to embed the emotional loading of some unit of language in a common emotion embedding space. This broadens the range of emotional nuances said models can capture. Importantly, our method learns a representation not for a specific unit of language itself but the emotion attached to it. This differs from previous work aiming to increase the affective load of, e.g., word embeddings (see below).

Emotion Embeddings. Several existing studies have used the term "emotion embeddings" (or similar phrasing) to characterize their work, yet either use the term in a different way or tackle a different problem compared to our study.

In more detail, Wang et al. (2020) present a method for increasing the emotional content of word embeddings based on re-ordering vectors according to the similarity in their emotion values, referring to the result as "emotional embeddings". Similarly, Xu et al. (2018) learn word embeddings that are particularly rich in affective information by sharing an embedding layer between models for different emotion-related tasks. They refer to these embeddings as "generalized emotion representation". Different from our work, these two studies primarily learn to represent *words* (with a focus on their affective meaning though), not emotions themselves. They are thus in line with previous research aiming to increase the affective load of word embeddings (Faruqui et al., 2015; Yu et al., 2017; Khosla et al., 2018).

Shantala et al. (2018) improve a dialogue system by augmenting their training data with emotion predictions from a separate system. Predicted emotion labels are fed into the dialogue model using a representation ("emotion embeddings") learned in a supervised fashion with the remainder of the model parameters. These embeddings are specific to their architecture and training dataset, they do not generalize to other label formats. Gaonkar et al. (2020) as well as Wang and Zong (2021) learn vector representations for emotion classes from annotated text datasets to explicitly model their semantics and inter-relatedness. Yet again, these emotion embeddings (the class representations) do not generalize to other datasets and label formats. Han et al. (2021) propose a framework for learning a common embedding space as a means of joining information from different modalities in multimodal emotion data. While these embeddings generalize over different modalities (audio and video), they do not generalize across languages and label formats. In summary, different from these studies, our emotion embeddings are not bound to any particular model architecture or dataset but instead generalize across domains and label formats, thus allowing to directly compare, say, English language items with BE5 ratings to Mandarin ones with VA ratings (see Tab. 1 vs. Fig. 1).

Coping with Incompatibility. In face of the variety of emotion formats, Felbo et al. (2017) present a transfer learning approach in which they pre-train a model with self-supervision to predict emojis in a large Twitter dataset, thus learning a representation that captures even subtle emotional nuances. Similarly, multi-task learning can be used to fit a model on multiple datasets potentially having different label formats, thus resulting in shared hidden representations (Tafreshi and Diab, 2018; Augenstein et al., 2018). While representations learned with these approaches generalize across different label formats, they do not generalize across model architectures or language domains.

Cross-lingual approaches learn a common latent representation for different languages but these representations are often specific to only one pair of languages and do not generalize to other label formats (Gao et al., 2015; Abdalla and Hirst, 2017; Barnes et al., 2018). Similarly, recent work with Multilingual BERT (Devlin et al., 2019) shows strong performance in cross-lingual zero-shot transfer (Lamprinidis et al., 2021), but samples from different languages still end up in different regions of the embedding space (Pires et al., 2019). These approaches are also specific to a particular model architecture so that they do not naturally carry over to, e.g., single-word emotion prediction. Multimodal approaches to emotion analysis show some similarity to our work, as they learn a common latent representation for several modalities which can be seen as separate domains (Zadeh et al., 2017; Han et al., 2021; Poria et al., 2019). However, these representations are typically specific to a single dataset and are not meant to generalize further.

In a recent survey on text emotion datasets, Bostan and Klinger (2018) point out naming inconsistencies between label formats. They build a joint resource that unifies twelve datasets under a common file format and annotation scheme. Annotations were unified based on the semantic closeness of their class names (e.g., merging *"happy"* and "*Joy*"). This approach is limited by its reliance on *manually* crafted rules which are difficult to formulate, especially for numerical label formats.

In contrast, emotion representation mapping (or "label mapping") aims at *automatically* learning such conversion schemes between formats from data (especially from "double-annotated" samples, such as the first two rows in Tab. 1; Stevenson et al., 2007; Calvo and Mac Kim, 2013; Buechel and Hahn, 2018a). As the name suggests, label mapping operates exclusively on the gold ratings, without actually deriving representations for language items. It can, however, be used as a post-processor, converting the prediction of another model to an alternative label format (used as a baseline in §4). Label mapping learns to transform *one* format *into another*, yet without establishing a more general representation. In a related study, De Bruyne et al. (2022) indeed do learn a common representation for different label formats by applying variational autoencoders to multiple emotion lexicons. However, their method still only operates exclusively on the gold ratings without actually predicting labels based on words or texts.

In summary, while there are methods to learn common emotion representations across *either* languages, linguistic domains, label formats, or model architectures, to the best of our knowledge, our proposal is the first to achieve all this simultaneously.

3 Methods

Let (X, Y) be a dataset with samples $X{:=}\{x_1, \dots x_n\}$ and labels $Y{:=}\{y_1, \dots, y_n\}$. The aim of emotion analysis is to find a model f that best predicts Y given X. Let us assume that the samples X are drawn from one of M domains $\mathcal{D}_1, \dots, \mathcal{D}_M$ and the labels are drawn from one of N label formats $\mathcal{L}_1, \dots, \mathcal{L}_N$. A domain refers to the vocabulary or a particular register of a given language (word- and text-level prediction). A label format is a set of valid labels with reference to particular emotion constructs. For instance, the VAD format consists of vectors (v, a, d) where the components v, a, d refer to *Valence*, *Arousal*, and *Dominance*, respectively, and are bound within a specified interval, e.g., $[1, 9]$.

3.1 Towards a Common Emotion Space

Fig. 2 provides an overview of our methodology. The naïve approach to emotion analysis is to learn separate models for each language domain, $\mathcal{D}_1, \dots, \mathcal{D}_M$, and label format, $\mathcal{L}_1, \dots, \mathcal{L}_N$, resulting in a potentially very high number of relatively weak models in terms of the emotional nuances they can capture *(a)*. The alternative we propose consists of two steps. First, we train a multi-way mapping that can translate between every pair of label formats $(\mathcal{L}_i, \mathcal{L}_j),\ i, j \in [1, N]$ via a shared intermediate representation layer, the common emotion space *(b)*. In a second step, we adopt existing model architectures to embed samples from a given domain in the emotion space, while the format-specific top layers of said mapping model are now utilized as portable prediction heads. The emotion space then acts as a mediating "interlingua" which connects each language domain, $\mathcal{D}_1, \dots, \mathcal{D}_M$, with each label format, $\mathcal{L}_1, \dots, \mathcal{L}_N$ *(c)*.

3.2 Prediction Head Training

A prediction head here refers to a function h that maps from a Euclidean input space $\mathbb{R}^d$ (the "emotion space") to a label format $\mathcal{L}_j$. We give prediction heads a purposefully minimalist design that consists only of a single linear layer without bias term. Thus, a head h predicts ratings $\hat{y}$ for an emotion embedding $x \in \mathbb{R}^d$ as $h(x) := Wx$, where W is a weight matrix. The reason for this simple head

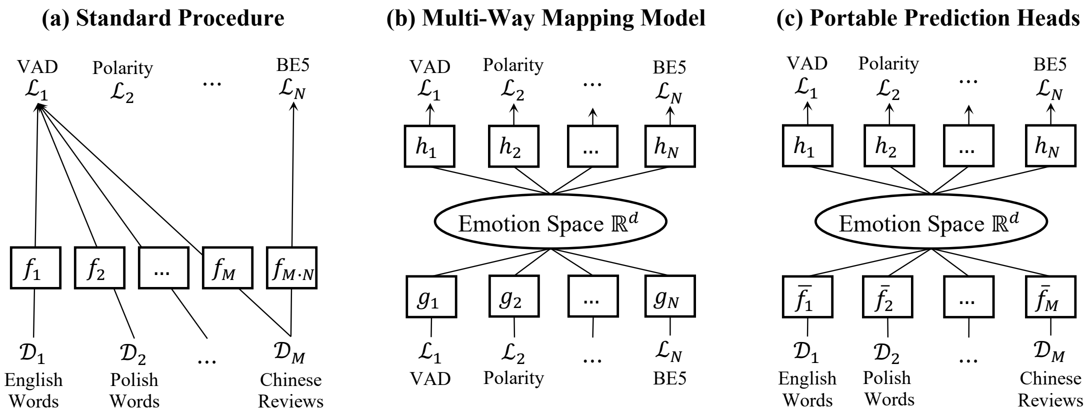


Figure 2: Overview of our methodology, illustrated by several choices of language domains and label formats.

design is to ensure that the affective information is more readily available in the emotion space. Alternatively, we can describe the weight matrix W as a concatenation of row vectors W_i, where each emotion variable corresponds to exactly one row. Thus, as a positive side effect of the lightweight design, we can directly locate emotion variables within the emotion space by interpreting their respective coefficients W_i as position vector (see Fig. 1).

Our challenge is to train a collection of heads $h_1, \dots h_N$ such that all heads produce *consistent* label outputs for a given emotion embedding from $\mathbb{R}^d$. For example, if the VAD head predicts a *joyful* VAD label, then the BE5 head should also produce a congruent *joyful* BE5 rating. In this sense, the prediction heads are "the heart and soul" of the emotion space: they define which affective state a region of the space corresponds to.

To devise a suitable training scheme for the heads, we first need to elaborate on our understanding of "consistency" between differently formatted emotion labels. We argue that an obvious case of such consistency is found in datasets for emotion label mapping (see §2). A label mapping dataset consists of two sets of labels following different formats $Y_1{:=}\{y_{1,1}, y_{1,2}, \dots y_{1,n}\}$ and $Y_2{:=}\{y_{2,1}, y_{2,2}, \dots y_{2,n}\}$, respectively. Typically, they are constructed by matching instances from independent annotation studies (e.g., the first two rows in Tab. 1). Thus, we can think of the two sets of labels as "translational equivalents", i.e., differently formatted emotion ratings, possibly capturing different affective nuances, yet still describing the same underlying expression of emotion in humans.

The intuition behind our training scheme is to "fuse" multiple mapping models by forcing them to produce the same intermediate representation for both mapping directions. This results in a multi-way mapping model with a shared representation layer in the middle (the common emotion space) followed by the prediction heads on top (Fig. 2*b*).

In more detail (see also Fig. 3 for an illustration of the following training procedure), let (Y_1, Y_2) be a mapping dataset with a sample (y_1, y_2). We introduce two new, auxiliary models g_1, g_2 that we call *label encoders*. Label encoders embed input ratings in the emotion space $\mathbb{R}^d$ and can be combined with the complementary prediction heads h_2, h_1 to form a mapping model (the subscript here refers to the label format). That is $h_2(g_1(y_1))$ yields predictions for y_2 and $h_1(g_2(y_2))$ for y_1.

Our goal is to align both the intermediate representations, $g_1(y_1)$, $g_2(y_2)$ while also deriving accurate mapping predictions. Therefore, we propose the following three training objectives:

$$
\begin{aligned}
L_{\text{map}} &:= \mathcal{C}[y_1, h_1(g_2(y_2))] + \mathcal{C}[y_2, h_2(g_1(y_1))] \\
L_{\text{auto}} &:= \mathcal{C}[y_1, h_1(g_1(y_1))] + \mathcal{C}[y_2, (h_2(g_2(y_2))] \\
L_{\text{sim}} &:= \mathcal{C}[g_1(y_1), g_2(y_2)]
\end{aligned}
$$

where $\mathcal{C}$ denotes the Mean-Squared-Error loss cri-

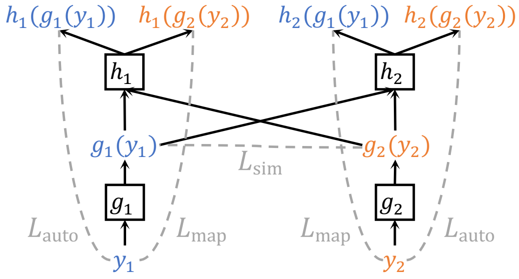


Figure 3: Training the Multi-Way Mapping Model.

terion. L_{map} is the *mapping loss* term where we compare true vs. predicted labels. The two summands represent the two mapping directions, assigning either of the two labels as the source, the other as the target format. The *autoencoder loss*, L_{auto}, captures how well the model can reconstruct the original input label from the hidden emotion representation. It is meant to supplement the mapping loss. Lastly, the *similarity loss*, L_{sim}, directly assesses whether both input label formats end up with a similar intermediate representation. The *total loss* for one instance, finally, is given by

$$L_{\text{total}} := L_{\text{map}} + L_{\text{auto}} + L_{\text{sim}}$$

In practice, we train a matching label encoder $g_1, \ldots, g_N$ for each of our prediction heads $h_1, \ldots, h_N$, thus covering all considered label formats $\mathcal{L}_1, \ldots \mathcal{L}_N$. All label encoders and prediction heads are trained simultaneously on a collection of mapping datasets. This is done as a hierarchical sampling procedure, where we first sample one of the mapping datasets (which determines the encoder and the head to be optimized in this step), then a randomly selected instance. The total loss is computed in a batch-wise fashion and the encoder and head parameters are updated via standard gradient descent-based techniques (see Appendix A for details). We use min-max scaling to normalize value ranges of the labels across datasets: for VAD we choose the interval $[-1, 1]$ and for BE5 the interval $[0, 1]$, reflecting their respective bipolar (VAD) and unipolar (BE5) nature (see Tab. 1).

3.3 Prediction Head Deployment

Following the training of the prediction heads $h_1, \ldots, h_N$, deploying them on top of a base model architecture f is relatively straightforward, resulting in a multi-headed model. The base model's output layer must be resized to the dimensionality of the emotion space $\mathbb{R}^d$ and any present nonlinearity (e.g, softmax or sigmoid activation) must be removed. This modified base model $\bar{f}$ is then optimized to produce emotion embeddings, the heads' input representation (see Fig. 4).

Head parameters are kept constant so that the base model is forced to optimize the representations it provides. Since the heads are specifically trained to treat emotion embeddings consistently, producing suitable representations for *one* head is also likely to produce suitable representations for the remaining heads. Yet, to avoid overfitting the base model to a particular one (i.e., producing representations that are particularly favorable for one head, but much less so for every other), each model $\bar{f}_i$ is trained using multiple heads depending on the available data.

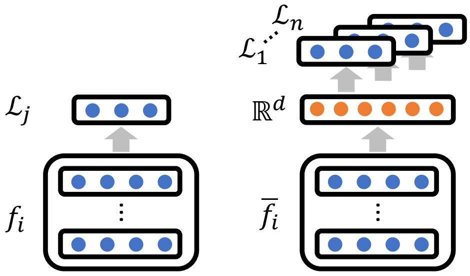


Figure 4: Schematic illustration of a base model before (left) and after (right) head deployment.

If *multiple* datasets are available that match the domain of the base model *and* use different label formats, we train the base model in a multi-task setup: We first draw one of the available datasets and then sample an instance (x, y) from there. Next, we derive a prediction using the matching head h_j as $\hat{y} := h_j(\bar{f}_i(x))$, before computing the *prediction loss*:

$$L_{\text{pred}} := \mathcal{C}[y, \hat{y}]$$

If, on the other hand, only *one* dataset is available which matches the domain of the base model $\bar{f}_i$, we complement the prediction loss with additional error signal using a newly proposed data augmentation technique. This method which we call *emotion label augmentation* synthesizes an alternative label $y^* := h_k(g_j(y))$ for a given instance (x, y) by taking advantage of the label encoder g_j that was trained in the previous step. While g_j translates the label y to the emotion space, the prediction head h_k provides labels in a format different from y. Those artificial labels are then used in place of actual gold labels resulting in the *data augmentation loss*

$$L_{\text{aug}} := \mathcal{C}[y^*, h_k(\bar{f}_i(x)]$$

where the second argument to the loss criterion $\mathcal{C}$ denotes the model's prediction for the previously synthesized labels. Then, $L_{\text{pred}} + L_{\text{aug}}$ yields the final loss.

4 Experimental Setup

The main idea behind our experimental setup is to compare a base model trained with the standard procedure against the same model with portable prediction heads (PPH) attached (cf. Fig. 2 *(a)* vs. *(c)*). Our goal is to show that we obtain the same, if not better, results using PPH compared with the naïve approach.

This study design reflects two purposes. First, comparing the base model with the PPH architecture yields experimental data that allow to indirectly assess the quality of the learned emotion representations. Second, such a comparison may help find evidence that the performance of the PPH approach *scales* with the employed base model—this would suggest that our method is likely to remain valuable even when today's state-of-the-art models are replaced by their successors. Importantly, we train only a single set of prediction heads. Thus, *all* experimental results of the PPH condition are based on the *same* underlying emotion space.

We distinguish two evaluation settings. In the first ("supervised") setting, train and test data come from (different parts of) the same dataset. Without PPH, we train one base model per dataset. Yet, with PPH, base models are shared across datasets of the same domain, whether or not their label formats agree. Consequently, the emotion space needs to store heterogeneous affective information in an easy-to-retrieve way (recall the "lightweight" head design; §3.2). Thus, positive evaluation results would indicate that our method learns a particularly rich representation of emotion. A practical advantage of PPH lies in the reduction of total disk space utilized by the resulting model checkpoints.

The second ("zero-shot") setting assumes that only *one* dataset per language is available, with one particular label format, but one would like to predict ratings in another format as well (e.g., imagine having a VA dataset for Mandarin but you are actually more interested in basic emotions for that language). Doing so with PPH is very simple—one only has to choose the desired head at inference time. Yet, doing so with the base model *per se* is simply impossible. To still be able to offer a quantitative comparison, we resort to an external label mapping component that translates the base model's output into the desired format. We emphasize that this is a very strong baseline due to the high accuracy of the label mapping approach, in general (Buechel and Hahn, 2018a). In this case, the practical advantage of the PPH approach lies in its independence of (possibly unavailable) external post-processors.

We conducted experiments on different word and text datasets. For words, we collected ten datasets (cf. Tab. 2) covering five languages. These data are structured as illustrated in the top half of Tab. 1. For text-level experiments we selected three corpora (cf. Tab. 3): Affective Text (AffT; Strapparava and Mihalcea, 2007), EmoBank (EmoB; Buechel and Hahn, 2017), and the Chinese Valence Arousal Texts (CVAT; Yu et al., 2016). For an illustration of the type and format of text-level data, see the bottom half in Tab. 1. Since these datasets comprise real-valued annotations, we will use Pearson Correlation r for measuring prediction quality. Datasets were partitioned into fixed train-dev-test splits with ratios ranging between 8-1-1 and 3-1-1; smaller datasets received larger dev and test shares.

The selected data govern how to train a given base model with PPH (§3.3). Since, except for Mandarin, there are always two datasets available per domain, we train the models in the supervised setting using the multi-task approach (but use emotion label augmentation for CVAT). By contrast, in the zero-shot setting, we train a model on *one*, yet test on *another* dataset. Thus, we rely on emotion label augmentation here (and have to exclude CVAT for a lack of a second Mandarin dataset). We emphasize that the zero-shot evaluation has very demanding data requirements: This setting not only requires two datasets of the *same* language domain with *different* label formats (which is already rare) but also additional data to fit mapping models for those particular label formats. To the best of our

ID	Vars	Size	Citation
en1	VAD	1,034	Bradley and Lang (1999)
en2	BE5	1,034	Stevenson et al. (2007)
es1	VA	14,031	Stadthagen-González et al. (2017)
es2	BE5	10,491	Stadthagen-González et al. (2018)
de1	VA	2,902	Võ et al. (2009)
de2	BE5	1,958	Briesemeister et al. (2011)
pl1	VA	2,902	Riegel et al. (2015)
pl2	BE5	2,902	Wierzba et al. (2015)
tr1	VA	2,029	Kapucu et al. (2018)
tr2	BE5	2,029	Kapucu et al. (2018)

Table 2: Word datasets. IDs contain the respective ISO 639-1 language code.

ID	Vars	Size	Lg	Domain
AffT	BE5	1,250	en	news headlines
EmoB	VAD	10,062	en	genre-balanced
CVAT	VA	2,969	zh	mixed online domains

Table 3: Overview of text datasets.

knowledge, EMOBANK and AFFT form the only suitable dataset pair on the text-level. At the word-level, such pairs are somewhat easier to get due to highly standardized data collection efforts for affective word norm datasets in psychology (see §2). For this reason, we employ a larger number of word- than text-level datasets in our experiments.

Importantly, only the data requirements for *evaluating* our approach in the zero-shot setting are hard to meet. Yet, *inference* is much easier to provide. We would even argue that the reason why our method is so hard to evaluate is precisely what makes it so valuable. Take the Mandarin CVAT dataset, for example. It is annotated with *Valence* and *Arousal*, but there is, to our knowledge, no compatible Mandarin dataset with basic emotions (thus, CVAT is not used in the zero-shot setting). Our method allows to freely switch between output label formats at inference time without language constraints. That is, we can predict BE5 ratings in Chinese even though there is no such training data.

In terms of base models, we used the Feed-Forward Network developed by Buechel and Hahn (2018b) for the word datasets. This model predicts emotion ratings based on pre-trained embedding vectors (taken from Grave et al., 2018). For text datasets, we chose the BERT$_{\text{base}}$ transformer model by Devlin et al. (2019) using the implementation and pre-trained weights by Wolf et al. (2020). Both (word and text) base models use identical hyperparameter settings with or without PPH extension. For the word model, we copied the settings of the authors, whereas text model hyperparameters were tuned manually for the base model *without* PPH.

We derived training data for the prediction heads (label mapping datasets) by combining the ratings of the word datasets $en1$ and $en2$. We used the label mapping model from Buechel and Hahn (2018a) as auxiliary label encoders. The dimensionality of the emotion space was set to 100. The label mapping models used as external post-processors in the zero-shot setting were also based on Buechel and Hahn (2018a) and were trained on the same data as the label encoders. Further details beneficial for reproducibility are given in Appendix D.

5 Results

Our main experimental results are summarized in Tables 4 to 7. For conciseness, correlation values are averaged over all target variables per dataset. Per-variable results are given in Appendix B.

Looking at the word datasets in the supervised setup (Tab. 4), we find that attaching portable prediction heads (PPH) not only retains, but often enough slightly increases the performance of the FFN base model (p=.008; two-sided Wilcoxon signed-rank test based on per-dataset results). Since we trained only one base model with PPH per language (but two without PPH), our data suggest that the emotion representations learned with PPH can easily hold affective information from different label formats at the same time. Moreover, PPH here offers the practical benefit of reducing the total disk space used by the resulting model checkpoints due to the smaller number of trained base models. Experiments on the text datasets using BERT as base model show results in line with these findings (see Tab. 5).

In the zero-shot setup, models are tested on datasets with label formats different from the training phase (e.g., $en1$ and $en2$). On the word datasets, using PPH shows small improvements in comparison with the base model as is (p=.003; Tab. 6), again suggesting that the learned emotion representations generalize robustly across label formats. Importantly, the base model is only capable of producing this label format *at all* because we equip it with a label mapping post-processor. While this procedure is very accurate (indeed, it constitutes a very strong baseline), it depends on an external component that may or may not be available for

	Base Model (FFN)		Base Model + PPH	
Test Data	Train Data	r	Train Data	r
en1 (VAD)	en1 (VAD)	.818	en1+en2	.824
en2 (BE5)	en2 (BE5)	.898	en1+en2	.898
es1 (VA)	es1 (VA)	.820	es1+es2	.833
es2 (BE5)	es2 (BE5)	.789	es1+es2	.820
de1 (VA)	de1 (VA)	.822	de1+de2	.836
de2 (BE5)	de2 (BE5)	.754	de1+de2	.748
pl1 (VA)	pl1 (VA)	.794	pl1+pl2	.835
pl2 (BE5)	pl2 (BE5)	.814	pl1+pl2	.845
tr1 (VA)	tr1 (VA)	.567	tr1+tr2	.575
tr2 (BE5)	tr2 (BE5)	.607	tr1+tr2	.614
Mean		.768		.783
Disk Use		4.33 MB		2.52 MB

Table 4: Word-level results of supervised setting.

	Base Model (BERT)		Base Model + PPH	
Test Data	Train Data	r	Train Data	r
EmoB	EmoB	.630	EmoB+AffT	.619
AffT	AffT	.746	EmoB+AffT	.755
CVAT	CVAT	.737	CVAT	.748
Mean		.704		.707
Disk Use		1.25 GB		0.81 GB

Table 5: Text-level results of supervised setting.

Test Data	Base Model (FFN) Train Data	r	Base Model + PPH Train Data	r
en1 (VAD)	en2 (BE5)	.801	en2	.810
en2 (BE5)	en1 (VAD)	.834	en1	.839
es1 (VA)	es2 (BE5)	.720	es2	.723
es2 (BE5)	es1 (VA)	.777	es1	.792
de1 (VA)	de2 (BE5)	.681	de2	.684
de2 (BE5)	de1 (VA)	.637	de1	.641
pl1 (VA)	pl2 (BE5)	.812	pl2	.812
pl2 (BE5)	pl1 (VA)	.787	pl1	.807
tr1 (VA)	tr2 (BE5)	.538	tr2	.563
tr2 (BE5)	tr1 (VA)	.550	tr1	.554
Mean		.714		.723
Method	ext. post-processor		built-in	

Table 6: Word-level results of zero-shot setting.

Test Data	Base Model (BERT) Train Data	r	Base Model + PPH Train Data	r
EmoB	AffT	.385	AffT	.407
AffT	EmoB	.584	EmoB	.582
Mean		.485		.495
Method	ext. post-processor		built-in	

Table 7: Text-level results of zero-shot setting.

the desired mapping direction (the source and the target label format). In contrast, the zero-shot capability is *innate* to ("built-in") the PPH approach. While we need only one prediction head per label format, the number of required mapping components for the base model grows on a quadratic scale with the number of considered formats. Again, text-level experiments show consistent results with word-level ones (Tab. 7).

One may object that the reduction of memory footprint shown in Tables 4 and 5 can also be achieved by traditional multi-task learning (i.e., attaching multiple heads to the base model, training it on two datasets, at once). Likewise, as Tables 6 and 7 indicate, the zero-shot capabilities offered by PPH can, in principle, be provided by additional label mapping components. However, PPH offers a much more elegant solution to combine the advantages of multi-task learning and label mapping without calling for additional (language) resources. Most importantly though, PPH is unique in its ability to embed samples from such heterogeneous datasets in a common representation space—a trait that may offer a general solution to studying emotion across languages, cultures, and individually preferred psychological theory.

6 Visualization of the Emotion Space

To gain first insights into the structure of our learned emotion space, we submitted the weight vectors of the emotion variables to principal component analysis (PCA; recall from §3.2 that each row in a head's weights matrix W corresponds to exactly one variable). Further, we derived emotion embeddings for the samples in Tab. 1 using the PPH-extended models evaluated in the last section. Applying the same PCA transformation to the embedding vectors, we co-locate the samples next to the emotion variables. The results (for the first three PCs) are displayed in Fig. 1. As can be seen, the relative positioning of the samples and variables shows high face validity—samples associated with similar feelings appear close to each other as well as to their akin variable. Appendix C provides additional analyses of the learned embedding space (focusing more deeply on the emotional interpretation of the PC axes and the distribution of emotion embeddings across languages) that further support this positive impression.

7 Conclusions & Future Work

We presented a method for learning a common representation space for the emotional loading of heterogeneous language items. While previous work successfully unified *some* sources' heterogeneity, our emotion embeddings are the first to *comprehensively generalize* over arbitrarily disparate language domains, label formats, and distinct neural network architectures. Our technique is based on a collection of *portable prediction heads* that can be attached to existing state-of-the-art models. Consequently, a model learns to *embed* language items in the common learned emotion space and thus to predict a wider range of emotional meaning facets, yet without sacrificing any predictive power as our experiments on 13 datasets (6 languages) indicate.

Since the resulting emotion representations both generalize across various use cases *and* evidently capture a rich set of affective nuances, we consider this work particularly useful for downstream applications. Thus, future work may build on a concept of *emotion similarity* to, e.g., cluster diverse language items by their associated feeling, retrieve words that evoke emotions similar to a query, or compare the affective meaning of phrases and concepts across cultures.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions and comments, and Tinghui Duan, doctoral student at the JULIE LAB, for assisting us with the Mandarin gold data.

References

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *IJCNLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 1: Long Papers, pages 506–515, Taipei, Taiwan, November 27 – December 1, 2017.

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 718–728, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *HLT-EMNLP 2005 — Proceedings of the Human Language Technology Conference & 2005 Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October 6–8, 2005.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *TSD 2007 — Proceedings of the 10th International Conference on Text, Speech and Dialogue*, pages 196–205, Pilsen, Czech Republic, September 3-7, 2007.

Silvio Amir, Ramón F. Astudillo, Wang Ling, Bruno Martins, Mário J. Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale Twitter sentiment lexicon induction. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*, pages 613–618, Denver, Colorado, USA, June 4-5, 2015.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1: Long Papers, pages 1896–1906, New Orleans, Louisiana, USA, June 1–6, 2018.

A. Balahur, J. M. Hermida, and A. Montoyo. 2012. Building and exploiting EMOTINET, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1):88–101.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2483–2493, Melbourne, Victoria, Australia, July 15–20, 2018.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GOODNEWSEVERYONE: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 1554–1566, Marseille, France, May 11–16, 2020.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August 20–26, 2018.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA.

Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. Discrete Emotion Norms for Nouns: Berlin Affective Word List (DENN–BAWL). *Behavior Research Methods*, 43(2):#441.

Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 578–585, Valencia, Spain, April 3–7, 2017.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904, Santa Fe, New Mexico, USA, August 20–26, 2018.

Sven Buechel and Udo Hahn. 2018b. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1: Long Papers, pages 1907–1918, New Orleans, Louisiana, USA, June 1–6, 2018.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2022. Joint emotion label space modeling for affect lexica. *Computer Speech & Language*, 71:#101257.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2020. An emotional mess! Deciding on a framework for building a Dutch emotion-annotated corpus. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 1643–1651, Marseille, France, May 11–16, 2020.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Challenge-HML 2020 — Proceedings of the 2nd Grand Challenge and Workshop on Multimodal Language @ ACL 2020*, pages 1–7, Virtual event, July 10, 2020.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *ACL 2020 — Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Virtual event, July 5–10, 2020.

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 — 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, USA, June 2–7, 2019.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL-HLT 2015 — Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, USA, May 31 – June 5, 2015.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark, September 7–11, 2017.

Dehong Gao, Furu Wei, Wenjie Li, Xiaohua Liu, and Ming Zhou. 2015. Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. *Computational Linguistics*, 41(1):21–40.

Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers. 2020. Modeling label semantics for predicting emotional reactions. In *ACL 2020 — Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4687–4692, Virtual event, Juli 5–10, 2020.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3483–3487, Miyazaki, Japan, May 7–12, 2018.

Jing Han, Zixing Zhang, Zhao Ren, and Björn W. Schuller. 2021. EmoBed : Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *IEEE Transactions on Affective Computing*, 12:553–564.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997 — Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July 7–12, 1997.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *COLING 2020 — Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Virtual Event, December 8–13, 2020.

Aycan Kapucu, Aslı Kılıç, Yıldız Özkılıç, and Bengisu Sarıbaz. 2018. Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological Reports*, pages 1–22. [Available online Dec 4, 2018].

Sopan Khosla, Niyati Chhaya, and Kushal Chawla. 2018. Aff2Vec : Affect–enriched distributional word representations. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2204–2218, Santa Fe, New Mexico, USA, August 20–26, 2018.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar, October 25–29, 2014.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal Joy : A data set and results for classifying emotions across languages. In *WASSA 2021 — Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EACL 2021*, pages 62–75, Virtual Event, April 19, 2021.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Shoushan Li, Jian Xu, Dong Zhang, and Guodong Zhou. 2016. Two-view label propagation to semi-supervised reader emotion classification. In *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics*, volume Technical Papers, pages 2647–2655, Osaka, Japan, December 11-16, 2016.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 174–184, Melbourne, Victoria, Australia, July 15–20, 2018.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-CANADA: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval 2013 — Proceedings of the 7th International Workshop on Semantic Evaluation @ NAACL-HLT 2013*, pages 321–327, Atlanta, Georgia, USA, June 14-15, 2013.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 28 – August 2, 2019.

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD : A multimodal multi-party dataset for emotion recognition in conversations. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 28 – August 2, 2019.

Ashequl Qadir and Ellen Riloff. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1209, Doha, Qatar, October 25–29, 2014.

Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL–R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California, USA, June 12–17, 2016.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Klaus R. Scherer. 2000. Psychological models of emotion. In Joan C. Borod, editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle H. Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 564–571, Valencia, Spain, April 3–7, 2017.

Roman Shantala, Gennadiv Kyselov, and Anna Kyselova. 2018. Neural dialogue system with emotion embeddings. In *SAIC 2018 — Proceedings of the 1st IEEE International Conference on System Analysis & Intelligent Computing*, pages 1–4, Kyiv, Ukraine, October 8–12, 2018.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 18-21, 2013.

Hans Stadthagen-González, Pilar Ferré, Miguel A. Pérez-Sánchez, Constance Imbault, and José Antonio Hinojosa. 2018. Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness. *Behavior Research Methods*, 50(5):1943–1952.

Hans Stadthagen-González, Constance Imbault, Miguel A. Pérez-Sánchez, and Marc Brysbært. 2017. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1):111–123.

Jacopo Staiano and Marco Guerini. 2014. DEPECHE MOOD: A lexicon for emotion analysis from crowd annotated news. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers, pages 427–433, Baltimore, Maryland, USA, June 22-27, 2014.

Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James. 2007. Characterization of the Affective

Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval 2007 Task 14: Affective text. In *SemEval 2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007*, pages 70–74, Prague, Czech Republic, June 23–24, 2007.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Shabnam Tafreshi and Mona Diab. 2018. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, volume 1, Technical Papers, pages 2905–2913, Santa Fe, New Mexico, USA, August 20–26, 2018.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy, July 28 – August 2, 2019.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Melissa L.-H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin Affective Word List Reloaded (BAWL–R). *Behavior Research Methods*, 41(2):534–538.

Shuo Wang, Aishan Maoliniyazi, Xinle Wu, and Xiaofeng Meng. 2020. EMO2VEC: Learning emotional embeddings via multi-emotion category. *ACM Transactions on Internet Technology*, 20(2):1–17.

Xiangyu Wang and Chengqing Zong. 2021. Distributed representations of emotion categories in emotion space. In *ACL-IJCNLP 2021 — Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & 11th International Joint Conference on Natural Language Processing*, volume 1: Long Papers, pages 2364–2375, Virtual Event, August 1-6, 2021.

Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLoS ONE*, 10(7):#e0132305.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771 [cs]*.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. EMO2VEC : Learning generalized emotion representation by multi-task training. In *WASSA 2018 — Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2018*, pages 292–298, Brussels, Belgium, October 31, 2018.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California, USA, June 12–17, 2016.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark, September 9–11, 2017.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, September 7–11, 2017.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing*, pages 165–176, Hong Kong, China, November 3-7, 2019.

Algorithm 1 Training the Multi-Way Mapping Model

```
1: $(Y_{1,1}, Y_{1,2}), (Y_{2,1}, Y_{2,2}), \dots (Y_{n,1}, Y_{n,2}) \leftarrow$ Mapping datasets used for training
2: $g_{1,1}, h_{1,1}, g_{1,2}, h_{1,2}, \dots, g_{n,1}, h_{n,1}, g_{n,2}, h_{n,2} \leftarrow$ randomly initialized label encoders and prediction heads $^\dagger$
3: $n_{\text{steps}} \leftarrow$ total number of training steps
4: for all $i_{\text{step}}$ in $1, \dots, n_{\text{steps}}$ do
5:     $(Y_{i,1}, Y_{i,2}) \leftarrow$ randomly sample a mapping dataset
6:     $(y_1, y_2) \leftarrow$ randomly sample a batch s.t. $y_1 \subset Y_{i,1}$ and $y_2 \subset Y_{i,2}$ with identical indices
7:     $(e_1, e_2) \leftarrow (g_{i,1}(y_1), g_{i,2}(y_2))$
8:     $\hat{y}_{1,1} \leftarrow h_{i,1}(e_1)$
9:     $\hat{y}_{1,2} \leftarrow h_{i,2}(e_1)$
10:    $\hat{y}_{2,1} \leftarrow h_{i,1}(e_2)$
11:    $\hat{y}_{2,2} \leftarrow h_{i,2}(e_2)$
12:    $L_{\text{map}} \leftarrow \mathcal{C}(y_1, \hat{y}_{2,1}) + \mathcal{C}(y_2, \hat{y}_{1,2})$ $^\ddagger$
13:    $L_{\text{auto}} \leftarrow \mathcal{C}(y_1, \hat{y}_{1,1}) + \mathcal{C}(y_2, \hat{y}_{2,2})$
14:    $L_{\text{sim}} \leftarrow \mathcal{C}(e_1, e_2)$
15:    $L_{\text{total}} \leftarrow L_{\text{map}} + L_{\text{auto}} + L_{\text{sim}}$
16:    compute $\nabla L_{\text{total}}$ and update weights
17: end for
```

[†] If two sets of labels $Y_{a,b}, Y_{c,d}$ follow the same label format, then they use the same label encoders (i.e, $g_{a,b} = g_{c,d}$) and prediction heads ($h_{a,b} = h_{c,d}$).

[‡] $\mathcal{C}$ denotes Mean-Squared-Error Loss.

A Algorithmic Details for Training the Multi-Way Mapping Model

The intuition behind Algorithm 1 is as follows: We simultaneously train multiple label encoders and prediction heads on several mapping datasets using three distinct objective functions. First, of course, we consider the quality of the label mapping (*mapping loss*; line 12). Second, we propose an *autoencoder loss* (line 13) where the model must learn to reconstruct the original input from the emotion embedding. Third, we propose an *embedding similarity loss* (line 14) which enforces the similarity of the hidden representation of both formats for a given instance since they supposedly describe the same emotion. Our training loop starts by first sampling one of the mapping datasets and then a batch from the chosen dataset (lines 5–6). To compute the loss efficiently, we first cache the encoded representations of both label formats (line 7) before applying all relevant prediction heads (lines 8–11).

B Per-Variable Results

For readability reasons, the experimental results reported in §5 only give the average performance score over all emotional target variables for a given dataset. To complement this, the full set of per-variable results are given in Tab. 8.

					Val	Aro	Dom	Joy	Ang	Sad	Fea	Dis	Mean
Level	Test	Setting	Model	Train									
word	en1	supervised	FFN	en1	.920	.704	.829	—	—	—	—	—	.818
			FFN+PPH	en1+en2	.936	.700	.836	—	—	—	—	—	.824
		zeroshot	FFN	en2	.932	.664	.808	—	—	—	—	—	.801
			FFN+PPH	en2	.927	.701	.802	—	—	—	—	—	.810
	en2	supervised	FFN	en2	—	—	—	.929	.900	.898	.890	.873	.898
			FFN+PPH	en1+en2	—	—	—	.936	.890	.895	.901	.869	.898
		zeroshot	FFN	en1	—	—	—	.918	.822	.805	.864	.759	.834
			FFN+PPH	en1	—	—	—	.914	.835	.850	.843	.751	.839
	es1	supervised	FFN	es2	.848	.792	—	—	—	—	—	—	.820
			FFN+PPH	es1+es2	.870	.795	—	—	—	—	—	—	.833
		zeroshot	FFN	es2	.873	.567	—	—	—	—	—	—	.720
			FFN+PPH	es2	.872	.575	—	—	—	—	—	—	.723
	es2	supervised	FFN	es2	—	—	—	.768	.793	.834	.803	.745	.789
			FFN+PPH	es1+es2	—	—	—	.817	.832	.857	.838	.754	.820
		zeroshot	FFN	es1	—	—	—	.808	.795	.823	.775	.685	.777
			FFN+PPH	es2	—	—	—	.811	.805	.839	.810	.695	.792
	de1	supervised	FFN	de1	.867	.776	—	—	—	—	—	—	.822
			FFN+PPH	de1+de2	.892	.780	—	—	—	—	—	—	.836
		zeroshot	FFN	de2	.832	.530	—	—	—	—	—	—	.681
			FFN+PPH	de2	.836	.532	—	—	—	—	—	—	.684
	de2	supervised	FFN	de2	—	—	—	.812	.766	.738	.798	.653	.754
			FFN+PPH	de1+de2	—	—	—	.842	.788	.655	.795	.662	.748
		zeroshot	FFN	de1	—	—	—	.824	.717	.500	.733	.411	.637
			FFN+PPH	de1	—	—	—	.824	.720	.489	.749	.424	.641
	pl1	supervised	FFN	pl1	.852	.735	—	—	—	—	—	—	.794
			FFN+PPH	pl1+pl2	.907	.764	—	—	—	—	—	—	.835
		zeroshot	FFN	pl2	.919	.705	—	—	—	—	—	—	.812
			FFN+PPH	pl2	.918	.707	—	—	—	—	—	—	.812
	pl2	supervised	FFN	pl2	—	—	—	.819	.807	.815	.810	.821	.814
			FFN+PPH	pl1+pl2	—	—	—	.897	.835	.820	.826	.846	.845
		zeroshot	FFN	pl1	—	—	—	.877	.786	.749	.763	.761	.787
			FFN+PPH	pl1	—	—	—	.893	.798	.777	.779	.789	.807
	tr1	supervised	FFN	tr1	.556	.577	—	—	—	—	—	—	.567
			FFN+PPH	tr1+tr2	.571	.579	—	—	—	—	—	—	.575
		zeroshot	FFN	tr2	.561	.514	—	—	—	—	—	—	.538
			FFN+PPH	tr2	.576	.549	—	—	—	—	—	—	.563
	tr2	supervised	FFN	tr1	—	—	—	.607	.603	.628	.627	.568	.607
			FFN+PPH	tr1+tr2	—	—	—	.611	.608	.628	.634	.589	.614
		zeroshot	FFN	tr1	—	—	—	.547	.566	.563	.579	.495	.550
			FFN+PPH	tr1	—	—	—	.583	.533	.575	.588	.488	.554
text	EmoB	supervised	BERT	EmoB	.801	.562	.527	—	—	—	—	—	.630
			BERT+PPH	EmoB+AffT	.798	.550	.509	—	—	—	—	—	.619
		zeroshot	BERT	AffT	.660	.200	.295	—	—	—	—	—	.385
			BERT+PPH	AffT	.686	.238	.297	—	—	—	—	—	.407
	AffT	supervised	BERT	AffT	—	—	—	.730	.634	.818	.836	.712	.746
			BERT+PPH	EmoB+AffT	—	—	—	.776	.659	.823	.841	.675	.755
		zeroshot	BERT	EmoB	—	—	—	.727	.485	.727	.689	.290	.584
			BERT+PPH	EmoB	—	—	—	.724	.491	.736	.704	.255	.582
	CVAT	supervised	BERT	CVAT	.878	.596	—	—	—	—	—	—	.737
			BERT+PPH	CVAT	.878	.617	—	—	—	—	—	—	.748

Table 8: Full experimental results per dataset and target variable in Pearson's r. "Mean" column corresponds to data given in Tabs. 4, 5, 6, and 7.

C Further Analysis of the Emotion Space

Building on the PCA transformation described in §6, we illustrate the position of *all* emotion variables in Fig. 5.

Within the first three principal components, two major groups can be visually discerned: the negative basic emotions of *Sadness*, *Fear*, and *Anger* forming the first group, and *Joy* and the two affective dimensions of *Valence* and *Dominance* forming the second. Intuitively speaking, this stands to reason, as *Valence* and *Dominance* typically show a very high positive correlation in annotation studies. The same holds for *Valence* and *Joy*. Likewise, *Sadness*, *Fear*, and *Anger* usually correlate positively with each other. Yet, between these groups of variables, studies show a negative correlation (cf. studies listed in Tab. 2). Interestingly, these observations indicate that the first principal component of the emotion space may represent a *Polarity* axis.

The remaining two variables, *Disgust* and *Arousal*, position themselves relatively far from the aforementioned groups and opposite of each other in the second principal component. While it is less obvious what this component represents, it is worth noting that both *Arousal* and *Disgust* generalize poorly across label formats. That is, while *Joy*, *Anger*, *Sadness*, and *Fear* are relatively easy to predict from VAD ratings in a label mapping experiment, and, likewise, *Valence* and *Dominance* can well be estimated from BE5 ratings, the variables of *Arousal* and *Disgust* seem to carry information more specific to their respective label format (Buechel and Hahn, 2018a). In the light of these observations, it may not come as a surprise that these variables receive positions that demarcate them clearly from the remaining ones.

The third principal component seems to be linked to the intensity or action potential of a feeling. Here, *Arousal*, *Dominance*, and *Disgust* and, less pronounced, *Fear* and *Anger* score highly, while *Sadness* and *Joy* receive comparatively low values.

Next, we examine whether the learned representations are sufficiently language-agnostic, i.e., that samples with similar emotional load receive similar embeddings independent of their language domain. We derived emotion embeddings for all entries in all of our word datasets (cf. Tab. 2) using the base models with portable prediction heads from the "supervised" setting of our main experiments. Again building on the previously established PCA transformation, we plotted the position of these multilingual samples in 2D (see Fig. 6).

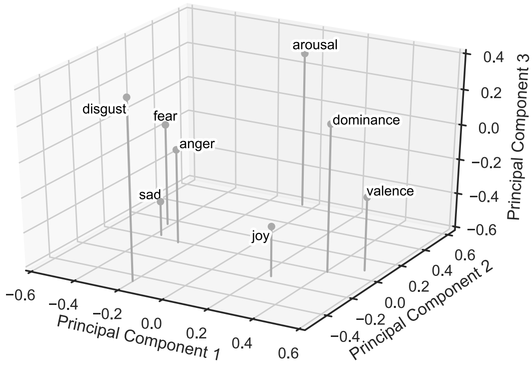


Figure 5: Position of emotion variables in PCA space.

It is noteworthy that entries in our emotion space seem to form clusters according to their affective meaning and not within their dataset or language. As a result, items from different languages overlap so heavily that their respective markers (○, △,□,◇, and ☆) become hard to differentiate.

Furthermore, we selected the highest- and lowest-rated words for *Valence* and *Arousal* and the highest-rated word for *Disgust* in each language. We locate these words in the PCA space and give translations for non-English entries. As can be seen, their position shows high face validity relative to each other and the emotion variables, supporting our claim that the learned emotion space is indeed language-independent.

We emphasize that monolingual, rather than crosslingual, word embeddings were used and that samples from each language were embedded using a separate base model. Hence, the observed alignment of words in PCA space may safely be attributed to our proposed training scheme using portable prediction heads.

D Further Details for Reproducibility

D.1 Description of Computing Infrastructure

All experiments were conducted on a single machine with a Debian 4 operating system. The hardware specifications are as follows:

- 1 GeForce GTX 1080 with 8 GB graphics memory
- 1 Intel i7 CPU with 3.60 GHz
- 64 GB RAM

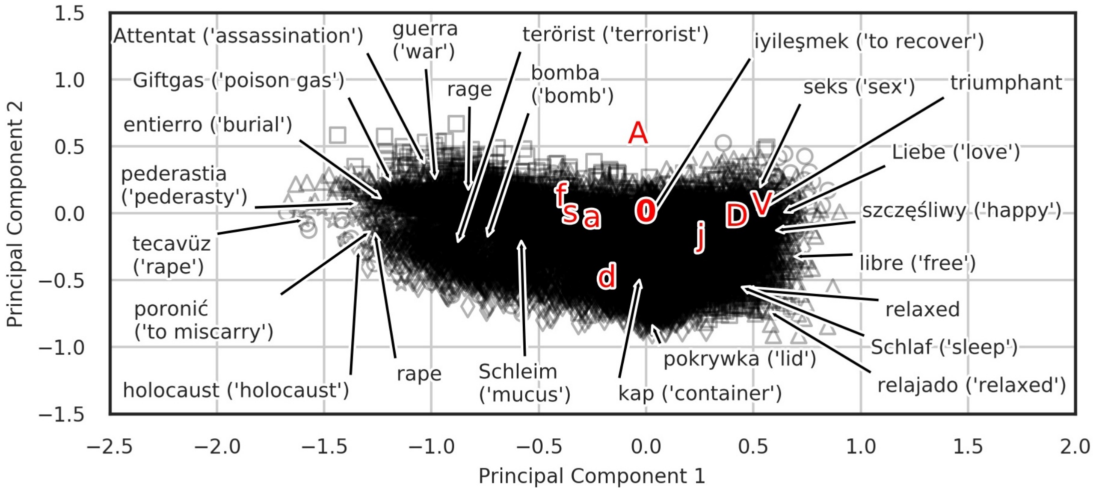


Figure 6: Position of the emotion variables **V***alence*, **A***rousal*, **D***ominance* and **j***oy*, **a***nger*, **s***adness*, **f***ear*, and **d***isgust* in the learned emotion space $\mathbb{R}^d$ (first two PCA dimensions; origin marked with "**0**") together with entries from English (◯), Spanish (△), German (□), Polish (◇), and Turkish (☆) word datasets, as well as highest and lowest *Valence* and *Arousal* word, and highest *Disgust* word per language (arrows).

D.2 Runtime of the Experiments

Training the multi-way mapping model takes about one minute. Training time for the base models varies depending on the dataset. In the following, we report training and inference times for the *largest* dataset per condition, respectively, describing an upper bound of the time requirements.

Regarding the word models, it takes about ten minutes to train a base model without portable prediction heads (PPH) and about 15 minutes to train one with PPH. Since the latter base model replaces two of the former ones in our experiments, the overall training time is reduced by using PPH. Training a word model with emotion label augmentation (the alternative technique for fitting a model with PPH) takes 10 minutes, about as long as training it without PPH. Inference is completed in 1.5 minutes in either case. However, most of that time is needed for loading the language-specific word embeddings. Once this task is done, actually computing the predictions takes only about one second.

Regarding the text models, a baseline model without PPH is trained in about 15 minutes. This number increases with PPH to 30 minutes using the multi-task approach (but again, one PPH model replaces two of the baseline models). In line with the runtime results of the word models, training the text base model with emotion label augmentation takes 15 minutes, about as long as training it without PPH. In either case, inference is completed in well under a minute.

D.3 Number of Parameters in Each Model

The number of parameters per model is given in Tab. 9.

Model (Component)	No. Parameters
Portable Prediction Heads	0.8K
Label Encoders (per format)	18.8K
Label Encoders (in total)	53.4K
Word-Level FFN (per model)	110.6K
BERT$_{\text{base}}$ (per model)	110.0M

Table 9: Number of parameters in each model.

D.4 Validation Performance

Tables 10 – 13 show the dev set results corresponding to the test set results in Tables 4 – 7, respectively. As can be seen, the former are consistent with the latter, yet overall slightly higher, as is usually the case.

D.5 Evaluation Metric

Prediction quality is evaluated using Pearson correlation defined as

$$r_{x,y} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $x = x_1, x_2, \ldots, x_n$, $y = y_1, y_2, \ldots, y_n$ are real-valued number sequences and $\bar{x}$, $\bar{y}$ are their

	Base Model (FFN)		**Base Model + PPH**	
Test Data	Train Data	r	Train Data	r
en1 (VAD)	en1 (VAD)	.800	en1+en2	.806
en2 (BE5)	en2 (BE5)	.876	en1+en2	.877
es1 (VA)	es1 (BE5)	.832	es1+es2	.850
es2 (BE5)	es2 (BE5)	.783	es1+es2	.820
de1 (VA)	de1 (BE5)	.825	de1+de2	.835
de2 (BE5)	de2 (BE5)	.780	de1+de2	.792
pl1 (VA)	pl1 (BE5)	.794	pl1+pl2	.841
pl2 (BE5)	pl2 (BE5)	.784	pl1+pl2	.835
tr1 (VA)	tr1 (BE5)	.600	tr1+tr2	.611
tr2 (BE5)	tr2 (BE5)	.613	tr1+tr2	.628
Mean		.769		.790
Disk Use		4.33 MB		2.52 MB

Table 10: Validation word-level results in the supervised setting.

	Base Model (BERT)		**Base Model + PPH**	
Test Data	Train Data	r	Train Data	r
EmoB	EmoB	.610	EmoB+AffT	.600
AffT	AffT	.783	EmoB+AffT	.790
CVAT	CVAT	.748	CVAT	.749
Mean		.714		.713
Disk Use		1.25 GB		0.81 GB

Table 11: Validation text-level results in the supervised setting.

	Base Model (FFN)		**Base Model + PPH**	
Test Data	Train Data	r	Train Data	r
en1 (VAD)	en2 (BE5)	.762	en2	.778
en2 (BE5)	en1 (VAD)	.814	en1	.815
es1 (VA)	es2 (BE5)	.759	es2	.758
es2 (BE5)	es1 (VA)	.767	es1	.779
de1 (VA)	de2 (BE5)	.692	de2	.672
de2 (BE5)	de1 (VA)	.696	de1	.696
pl1 (VA)	pl2 (BE5)	.806	pl2	.829
pl2 (BE5)	pl1 (VA)	.776	pl1	.796
tr1 (VA)	tr2 (BE5)	.556	tr2	.571
tr2 (BE5)	tr1 (VA)	.556	tr1	.565
Mean		.719		.726
Method		ext. post-processor		built-in

Table 12: Validation word-level results in the zero-shot setting.

	Base Model (BERT)		**Base Model + PPH**	
Test Data	Train Data	r	Train Data	r
EmoB	AffT	.353	AffT	.368
AffT	EmoB	.636	EmoB	.664
Mean		.495		.516
Method		ext. post-processor		built-in

Table 13: Validation text-level results in the zero-shot setting.

respective means. We rely on the implementation provided in the SciPy package.[2]

D.6 Model and Hyperparameter Selection

As described in §4, we mostly relied on hyperparameter choices by the authors of our base models. Hence, we performed only a relatively small amount of tuning throughout this work.

For the word base model and the label encoder, no further hyperparameter selection was required. For the text base model (BERT), we verified via a first round of development experiments that default settings yield satisfying prediction quality on our datasets. The learning rate of the AdamW optimizer was set to 10^{-5} based on established recommendations. Besides the number of training epochs (see below), the only dataset-specific hyperparameter choice had to be made for the batch size which we set according to constraints in GPU memory. (The samples in the CVAT dataset are significantly longer than in AffT so that fewer samples of the former can be placed in one batch.) We used the pre-trained weights "bert-base-uncased" and "bert-base-chinese" from Wolf et al. (2020) for the English and Mandarin datasets, respectively. The dimensionality of the emotion space $\mathbb{R}^d$ was initially set to 100 and remained unchanged after verifying that the Multi-Way Mapping Model indeed showed good label mapping performance.

For each (word or text) dataset, we trained the models well beyond convergence, recording their dev set performance after each epoch (number of epochs differs between datasets). We then chose the best-performing checkpoint (according to Pearson correlation) for the final test set evaluation.

Hyperparameter choices were identical between base models with and without PPH. We emphasize that for each base model, hyperparameters were set (by us or by the respective authors) with respect to base model *without* PPH, thus forming a challenging testbed for our approach. We see an extensive hyperparameter search as a fruitful venue for future work.

D.7 Data Access

Below, we list URLs for all datasets used in our experiments.

en1 https://osf.io/2k97q/download (ratings must be extracted from PDF)

en2 https://static-content.springer.com/esm/art%3A10.3758%2FBF03192999/MediaObjects/Stevenson-BRM-2007.zip

es1 https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-015-0700-2/MediaObjects/13428_2015_700_MOESM1_ESM.csv

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

es2 https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-017-0962-y/MediaObjects/13428_2017_962_MOESM1_ESM.csv

de1 https://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/allgpsy/Download/BAWL/index.html

de2 https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-011-0059-y/MediaObjects/13428_2011_59_MOESM1_ESM.xls

pl1 https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-014-0552-1/MediaObjects/13428_2014_552_MOESM1_ESM.xlsx

pl2 https://doi.org/10.1371/journal.pone.0132305.s004

tr1 https://osf.io/rxtdm

tr2 https://osf.io/rxtdm

AFFT http://web.eecs.umich.edu/~mihalcea/affectivetext/

EMOB https://github.com/JULIELab/EmoBank

CVAT http://nlp.innobic.yzu.edu.tw/resources/cvat.html

D.8 Details of Train-Dev-Test Splits

EMOB comes with a stratified split with ratios of about 8-1-1 (exactly 8062 train, 1000 dev, 1000 test samples). Since the samples of AFFT are mostly also included in EMOB, we decided to use the data split of the latter for the former, too. Samples of AFFT that were not included in EMOB (about 5% of the data) were removed before the experiments. CVAT features a 5-fold data split but without assigning the resulting parts to train, dev, or test utilization. We used the first three for training, the fourth for development/validation, and the fifth for testing.

The word datasets in Tab. 2 do not come with a fixed data split. Instead, we defined splits ourselves with ratios ranging between 3-1-1 to 8-1-1, depending on the number of samples. Instances were randomly assigned to train, dev, and test split using fixed random seeds. The resulting partitions were stored as JSON files and placed under version control.

Part III

Information on the Author

14 Full Publication List

Publications included in this thesis are set in **bold** font.

2021

Sven Buechel, Luise Modersohn, and Udo Hahn. 2021. Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104.

2020

Sven Buechel, João Sedoc, H. Andrew Schwartz, and Lyle Ungar. 2020b. Learning emotion from 100 observations: Unexpected robustness of deep learning under strong data limitations. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 129–139.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020a. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1664–1673.

2019

Sven Buechel, Simon Junker, Thore Schlaak, Claus Michelsen, and Udo Hahn. 2019. A time series analysis of emotional loading in central bank statements. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 16–21.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2019. Modeling word emotion in historical language: Quantity beats supposed stability in seed word selection. In *Proceedings of the Third Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11.

2018

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JeSemE: Interleaving semantics and emotions in a Web service for the exploration of language change phenomena. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 10–14.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904.

Maria Moritz, Johannes Hellrich, and Sven Büchel. 2018a. A method for human-interpretable paraphrasticality prediction. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 113–118.

Sebastian G.M. Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. 2018. A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31.

Maria Moritz, Johannes Hellrich, and Sven Buechel. 2018b. Towards a metric for paraphrastic modification. In *Digital Humanities 2018. Book of Abstracts*, pages 457–459.

Sven Buechel and Udo Hahn. 2018c. Word emotion induction for multiple languages as a deep multi-task learning problem. 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 1907–1918.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution — A case study to get around IPRs and privacy constraints featuring the German JSynCC corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 1259–1266.

Sven Buechel and Udo Hahn. 2018b. Representation mapping: A novel approach to generate high-quality multi-lingual emotion lexicons. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 184–191.

2017

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2017. The course of emotion in three centuries of German text—A methodological framework. In *Digitial Humanities 2017. Conference Abstracts*, pages 176–179.

Sven Buechel and Udo Hahn. 2017b. A flexible mapping scheme for discrete and dimensional emotion representations: Evidence from textual stimuli. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 180–185.

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Sven Buechel and Udo Hahn. 2017c. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12.

2016

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016b. Feelings from the past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities*, pages 54–61.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122.

Sven Buechel, Udo Hahn, Jan Goldenstein, Sebastian G. M. Händschke, and Peter Walgenbach. 2016a. Do enterprises have emotions? In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 147–153.

2015

Johannes Hellrich, Stefan Schulz, Sven Buechel, and Udo Hahn. 2015. JUFit: A configurable rule engine for filtering and generating new multilingual UMLS terms. In *AMIA Annual Symposium Proceedings 2015*, pages 604–610.

15 Declaration of Authorship

Hiermit erkläre ich ehrenwörtlich, dass:

- mir die geltende Promotionsordnung bekannt ist.
- ich die Dissertation selbst angefertigt und keine Textabschnitte eines anderen Autors oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen angegeben habe.
- mich nur die genannten Koautoren bzw. Korrekturleser (siehe Acknowledgments) bei der Auswahl und Auswertung des Materials und der Herstellung des Manuskripts unterstützt haben.
- ich nicht die Hilfe eines Promotionsberaters in Anspruch genommen habe und Dritte weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.
- ich die Dissertation noch nicht als Prüfungsarbeit für eine wissenschaftliche Prüfung eingereicht habe.
- ich nicht die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Jena, den ....

Sven Büchel