3-30-2022

# Detecting the Emotions of Animate Beings in Narrative

Samira Zad
*Florida International University*, szad001@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

# DETECTING THE EMOTIONS OF ANIMATE BEINGS IN NARRATIVE

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Samira Zad

2022

To: John L. Volakis
College of Engineering and Computing

This dissertation, written by Samira Zad, and entitled Detecting the Emotions of Animate Beings in Narrative, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Monique Ross

Armando Barreto

Fahad Saeed

Mark A. Finlayson, Major Professor

Date of Defense: March 11, 2022

The dissertation of Samira Zad is approved.

John L. Volakis
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
Dean of the University Graduate School

Florida International University, 2022

DEDICATION

To my mother Simin

for her endless love, support and encouragement

ACKNOWLEDGMENTS

I want to express my gratitude to my advisor, Dr. Mark A. Finlayson; without his encouragement and guidance, I would not have been able to complete my Ph.D. journey. He taught me a lot of skills and techniques in Natural Language Processing and Computer Science in general and it was a privilege to work in the Cognac lab under his supervision. I would also like to thank my committee members, Dr. Monique Ross, Dr. Armando Barreto, and Dr. Fahad Saeed. I appreciate the time and valuable advice I received from them throughout my Ph.D. I am grateful to Joshuan Jimenez, a graduate student whom I mentored throughout his undergraduate and graduate studies. I am also thankful to my colleagues and friends from Cognac Lab including Victor, Deya, Muhammed, Labiba, Anurag, Mustafa, Mireya, and Armando.

ABSTRACT OF THE DISSERTATION

DETECTING THE EMOTIONS OF ANIMATE BEINGS IN NARRATIVE

by

Samira Zad

Florida International University, 2022

Miami, Florida

Professor Mark A. Finlayson, Major Professor

Identifying emotions as expressed in text (a.k.a. text emotion recognition) has received a lot of attention over the past decade. Narratives often involve a great deal of emotional expression, and so emotion recognition on narrative text is of great interest to computational approaches to narrative understanding. The meaning and impact of narratives is strongly bound up with the emotions expressed therein. Emotions may be experienced by characters in a story (which may include the narrator), by a story-external narrator, or by the reader.

There has been so far two separate streams of work relevant to this observation: (1) emotion detection, and (2) detection of animate beings. These two streams have not yet been combined to attempt to identify the emotions experienced by animate beings in the text. In this dissertation, I use the two streams to construct a computational framework for detecting the emotions experienced by animate beings in a given text.

In the first step, I design a high-performing approach to emotion detection in narrative text and explore three techniques NMF, PCA, and LDA that NMF performed best, with an overall $F_1$ of 0.809.

The second step involves identifying and improving an emotion lexicon that will be used by my animate beings' emotion recognition system. I describe a procedure for semi-automatically correcting these problems in the NRC.

In the third step, I provide the ABBE corpus—Animate Beings Being Emotional—a new double-annotated corpus of texts. Plutchik's 8-category emotion model was used to categorize the emotion expressions, and the overall inter-annotator agreement for the annotations was 0.83 Cohen's Kappa ($kappa$), indicating excellent agreement.

Finally, I demonstrate an emotion detection system based on a non-neural machine learning classifier to identify the emotions expressed as being experienced by animate beings. I use Plutchik's emotion model, as well as the Revised NRC Emotion Lexicon. I train my model and evaluate my results using ABBE that has been annotated for animate beings, emotions, and the connections between them. The system achieves an overall micro $F_1$ of 0.76 when using gold-standard animate beings, and 0.60 when relying on computed animate beings.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

Chapter 1

# Introduction

## 1.1 Motivation

Natural Language Processing (NLP) is the branch of artificial intelligence that involves automating the preparation and analysis of large amounts of textual and verbal data. One of the main focuses of NLP is to make use of machine learning approaches to design and construct computational platforms. These platforms automate the process of extracting knowledge from both structured and unstructured sources to facilitate searching through large amounts of textual data in a short period of time and to obtain appropriate information. One of these important platforms is Text-Based Emotion Detection (TBED), which has been used by researchers to automatically detect affect, identifying the feelings and sentiments expressed in a text. It detects emotions from a variety of data sources using natural language processing, computational linguistics, and psychological emotion theories.

Emotion is a primary aspect of communication and can be transmitted across many modalities, including gestures, facial expressions, speech, and text [Perikos and Hatzilygeroudis, 2013]. Text-Based Emotion Detection (TBED), one of the fastest growing branches of Natural Language Processing (NLP), is the process of classifying syntactic or semantic units of a corpus into a given set of emotion classes proposed by a psychological model. Automatic Text-Based Emotion Detection mechanisms use machine learning approaches to create computational platforms automating the process of extracting emotions from both structured (e.g., books and articles) and unstructured text sources (like comments on social media). Text-Based Emotion Detection has a wide variety of applications in the area of artificial intelligence: obtaining insight into public opinion on various socio-political subjects to better understand public opinion and narratives, ex-

tracting characters' emotions expressed by the narrator of a story, Semantic analysis of documents and public messages related to terrorist attacks (to mitigate risks), automated analysis of historical corpora, study of product reviews (to assess customer satisfaction), and accurately analyzing collected data pertaining to the disaster situations and people who were affected by the disaster as part of Humanitarian Assistance and Disaster Relief (HADR) efforts are some examples.

As stated previously, Text-Based Emotion Detection is useful for many applications, including for automated narrative understanding. A narrative is "a representation of connected events and characters that has an identifiable structure, is bounded in space and time, and contains implicit or explicit messages about the topic being addressed" [Kreuter et al., 2007, p. 222], and narratives are often used to express the emotions of authors and characters, as well as induce emotions in audiences. For many narratives—one need only consider romances such as *Romeo and Juliet* or the movie *Titanic*—it is no exaggeration to say that lacking an understanding of emotion leads to a seriously impoverished view of the meaning of the narrative.

Text-Based Emotion Detection is a challenging problem on account of the complex relationship between felt emotion and linguistic expression. This includes not only standard natural language processing challenges, such as polysemous words and the difficulty of co-reference resolution [Uzuner et al., 2012, Peng et al., 2019], but also emotion-specific challenges such as how context can subtly change emotional interpretations [Cowie et al., 2005]. These technical challenges are exacerbated by a shortage of quality labeled data addressing this task.

Similarly, there exist recently developed approaches to detect animate beings. However, no one has yet integrated these techniques for detecting emotions expressed as being experienced by the animate beings in a narrative. This work is an example of such a system.

## 1.2 Broader Impact

Much work has been done in the field of sentiment analysis on online texts [Zad et al., 2021a]. However, recently, there is a high demand to explore and pay attention to *emotion* detection on texts [Zad et al., 2021b]. Emotion detection is an active research area within Natural Language Processing (NLP). The goal of emotion detection is to computationally extract and quantify emotional states expressed in a text, including narrative text specifically. My particular interest is in animate being emotion identification in narrative. Extracting emotion for identifying animate beings in narratives like Romeo and Juliet or Titanic are critical to computational understanding of narratives.

My dissertation is split into two components. The first phase is to construct an emotion detector. In the second phase, I adapt this system to the detection of emotional states associated with specific animate beings.

In the first step, I design a high performing approach to emotion recognition in narrative text and carefully implement and characterize the technique, exploring a design space of three different noise cancellation or dimension reduction techniques (NMF, PCA, or LDA), exploring various hyper-parameter settings. My experiments indicate that NMF performed best, with an overall $F_1$ of 0.809.

In the second step, I identify and improve an emotion lexicon to be used for my animate beings emotion detection system. There have been several attempts to create an accurate and thorough emotion lexicon in English, which identifies the emotional content of words. Of the several commonly used resources, the NRC emotion lexicon has received the most attention due to its availability, size, and its choice of Plutchik's expressive 8-class emotion model. In this work, I identify a large number of troubling entries in the NRC lexicon, where words that should in most contexts be emotionally neutral, with no affect (e.g., *lesbian*, *stone*, *mountain*), are associated with emotional labels that are

inaccurate, nonsensical, pejorative, or, at best, highly contingent and context-dependent (e.g., *lesbian* labeled as DISGUST and SADNESS, *stone* as ANGER, or *mountain* as ANTICIPATION). I describe a procedure for semi-automatically correcting these problems in the NRC, which includes disambiguating POS categories and aligning NRC entries with other emotion lexicons to infer the accuracy of labels. I demonstrate via an experimental benchmark that the quality of the resources is thus improved. Joshuan Jimenez, a graduate student in the Cognac lab, assisted me with the manual part.

In the third step, to develop my animate being emotion detection system, I and Joshuan Jimenez provide the ABBE corpus—Animate Beings Being Emotional—a new double-annotated corpus of texts that captures this key information for one class of emotion experiencer, namely, animate beings in the world described by the text. Such a corpus is useful for developing systems that seek to model or understand this specific type of expressed emotion. Our corpus contains 30 chapters, comprising 134,513 words, drawn from the Corpus of English Novels, and contains 2,010 unique emotion expressions attributable to 2,227 animate beings. The emotion expressions are categorized according to Plutchik's 8-category emotion model, and the overall inter-annotator agreement for the annotations was 0.83 Cohen's Kappa ($\kappa$), indicating excellent agreement.

Finally, I demonstrate an emotion detection system based on a non-neural machine learning classifier to identify the emotions expressed as being experienced by animate beings. I use Plutchik's emotion model (JOY, SADNESS, ANGER, FEAR, SURPRISE, ANTICIPATION, TRUST, and DISGUST), as well as the Revised NRC Emotion Lexicon developed in Step two. I train my model and evaluate my results using ABBE that has been annotated for animate beings, emotions, and the connections between them in the previous step.The system achieves an overall micro $F_1$ of 0.76 when using gold-standard animate beings, and 0.60 when relying on computed animate beings, showing that this task is more challenging than expected.

## 1.3 Phase 1: Sentence Emotion Detection

### 1.3.1 Approach

This component is made of four consecutive steps: In the first step, pre-processing, the system processes the input corpus using the CoreNLP library [Manning et al., 2014] to separate the text into sentences and lemmatize sentences to obtain tokens making the corpus. In the second step, vector space modeling, I used the lemmatized tokens to generate a vector representation of the emotional content of a sentence. In the third step, noise cancellation or dimension reduction, I explored three different models to either reduce dimensions or extract features of the vector space. One of the main contributions here was to analyze and explain the effect of this step on the performance of the final emotion recognition system. Finally, the fourth step, labeling, compared the vector for each sentence with vectors for each emotion, choosing the closest emotion as the label for the sentence.

### 1.3.2 Data

To implement the emotion sentence detection system, I began with a corpus of manually annotated fairy tales constructed by [Alm, 2008], comprising 176 children's fairy tales (80 from Brothers Grimm, 77 from Hans Andersen, and 19 from Beatrix Potter) with 15,087 unique sentences (15,302 sentences), 7,522 unique words and 320,521 total words. These fairy tales were annotated by two annotators labeling the emotion and mood of each sentence as one of joy, anger, fear, sadness, or neutral which resulted in four labels per sentence. Across the sentences, only 1,090 of them agreed on *all four non-neutral labels*.

I used the WordNet Affect [Strapparava and Valitutti, 2004], linguistic resource, which builds upon the general WordNet database [Fellbaum, 1998a] to associate specific words

with specific emotions. WNA classifies 280 WordNet *Noun* synsets into an emotion hierarchy rooted in an augmented version of Ekman's basic emotions. WordNet links an additional 1,191 *Verb*, *Adverb*, and *Adjective* synsets to this core *Noun*-focused hierarchy. These synsets represent approximately 3,500 English lemma-POS pairs.

### 1.3.3 Method and Results

The flowchart Figure 1.1 is a superset of the implemented emotion sentence identification system. In the pre-processing step, I constructed a bag of words for each sentence in the given corpus by tokenizing the sentence and lemmatizing each word. Then, I computed a *tf-idf* vector for each sentence as well as a standard vector for each emotion label (Step 2). For each sentence I constructed an $m$ dimensional vector where each entry in the vector is the *tf-idf* of an emotional term in the sentence. The constructed vector space model is represented by a matrix $V$. Also, I computed a standard vector for each emotion class in the same space, by using the WordNet Affect terms associated with label. In step three, the vectors $V_s$ and $Y_e$ from the previous step contains components related to many terms that have little or no effect on the emotion labeling of their sentences. For unsupervised learning, I used dimensional reduction or noise cancellation techniques to significantly improve the performance of the emotion detection process. I explored Principle Component Analysis (PCA) [Abdi and Williams, 2010], Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and Non-Negative Matrix Factorization (NMF) [Lee and Seung, 1999]. When using PCA or LDA, one can move directly to fourth step of the system; however, in the case of NMF, must select important terms, remove irrelevant features, and reconstruct the vector space (Step 3.1).

The emotion recognition process takes place by measuring the similarity between sentence vectors $V_s$ and standard emotion vectors $Y_e$ using cosine-similarity method. Ap-

plying this method on Fairy Tales corpus resulted in 80.9% $F_1$-score. This work was published in the 1st International Workshop on Narrative Understanding, Storylines, and Events (NUSE 2020) and was held concurrently with ACL'2020. Chapter §3 addresses this work in greater detail.

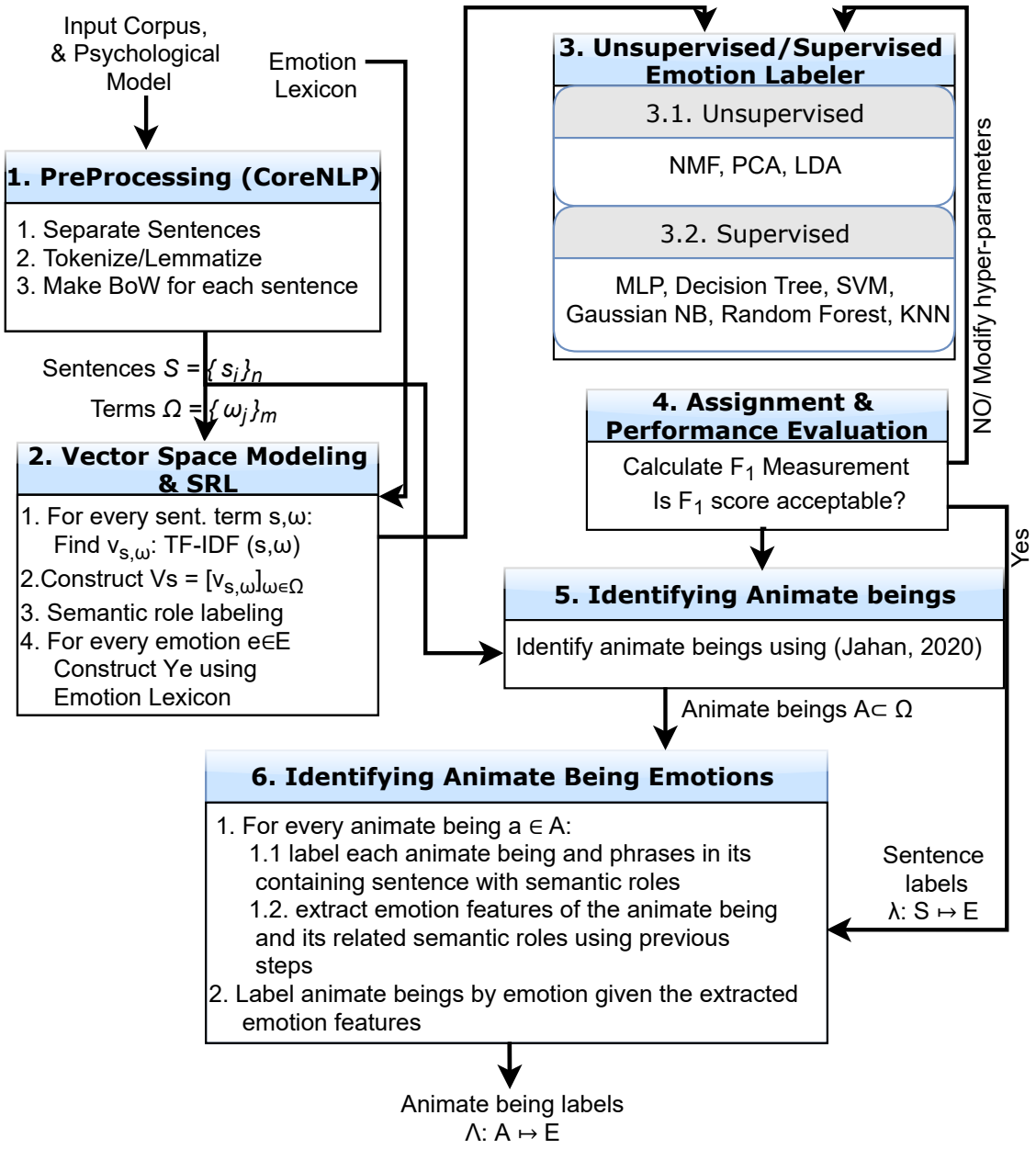



Figure 1.1: Flowchart of the proposed system.

## 1.4 Phase 2: Animate being Emotion Detection

The main phase of my dissertation is to build upon Phase 1 work (sentence emotion identification) to create a system which can identify the emotions of specific animate beings.

Animate beings are vital elements in narrative structures and identifying them are crucial for automatically understanding narratives [Jahan et al., 2018a]. Animate beings often have emotions, and these emotions drive or explain the action in the story. Therefore, identification of the emotions of animate beings can enhance the automation process of understanding narratives and give significant value to the prior work in this field.

### 1.4.1 Approach

The second phase comprise two consecutive steps: (1) identify animate beings of stories in an input corpus using state-of-the-art method designed and implemented by [Jahan et al., 2020a] for animate being identification and (Step 5 in Figure 1.1); and (2) label animate beings by emotions given their semantic-role labels as well as their emotion features extracted from emotion labels generated by modified form of the emotion identification system developed in Phase 1 (Step 6 in Figure 1.1).

#### Animate being Identification

This task is the fifth step of my proposed pipeline depicted in Figure 1.1. To identify animate beings of a given corpus, I use a classification model applied to the following four features mentioned in my labmate's work [Jahan et al., 2020a]: (a) *Coreference Chain Length (CL)*: the length of a coreference chain as an integer feature which explicitly captures the tendency of the long chains to be animate beings discussed in [Eisenberg and Finlayson, 2017]. (b) *Semantic Subject (SS)*: Binary identifier of whether the head of a

coreference chain appears as a semantic subject (ARG0) to a verb computed by the semantic role labeler associated with the Story Workbench annotation tool in [Finlayson, 2008, 2011]. (c) *Named Entity (NE)*: Binary identifier of whether the head of a coreference chain is a named entity with the category *PERSON* computed using the classic API of the Stanford dependency parse [Manning et al., 2014, v3.7.0]. (d) *WordNet (WN)*: Binary identifier of whether the head of a coreference chain is a descendant of *person* in WordNet.

**Emotion Extraction of Animate beings**

Using the detected animate being, I then developed a method to associate identified emotions with specific Animate beings. This task which is the sixth and last step of my proposed pipeline depicted in Figure 1.1, and is addressed in §6 in great detail. The proposed classification task is designed to extract the emotions of each animate being. For this work, I applied Plutchick's emotion model (JOY, SADNESS, ANGER, FEAR, SURPRISE, ANTICIPATION, TRUST, and DISGUST) [Plutchik, 1980, 1984, 1994] as well as the Revised NRC Emotion Lexicon [Zad et al., 2021c] (addressed in §4). Next, I used thirty chapters from the Corpus of English Novels (CEN) that had been annotated for animate beings, emotions, and their relations to train my model and evaluate my outcomes. Next, I used ABBE corpus—Animate Beings Being Emotional—a new double-annotated corpus of texts that captures this key information for one class of emotion experiencer, namely, animate beings in the world described by the text [Zad et al., 2022] that is addressed in Chapter §5 in detail, to train and evaluate Animate beings emotion detection model. The system achieves an overall micro $F_1$ of 0.76 when using gold-standard animate beings, and 0.60 when relying on computed animate beings, indicating that this task is more difficult than expected, is addressed in Chapter §6.

### 1.4.2 Data

For the sake of training and testing the system proposed in the second phase, I and Joshuan Jimenez had manually annotated 30 chapters of the Corpus of English Novels by applying Plutchik's psychological model [Plutchik, 1994] and assigning emotion labels to each animate being.

Emotion detection is an established NLP task of demonstrated utility for text understanding. However, basic emotion detection leaves out key information, namely, *who* is experiencing the emotion in question. For example, it may be the author, the narrator, or a character; or the emotion may correspond to something the audience is supposed to feel, or even be unattributable to a specific being, e.g., when emotions are being discussed *per se*. This work which has resulted in the ABBE corpus—Animate Beings Being Emotional—contains 30 chapters, comprising 134,513 words, drawn from the Corpus of English Novels, and contains 2,010 unique emotion expressions attributable to 2,227 animate beings [Zad et al., 2022]. The emotion expressions are categorized according to Plutchik's 8-category emotion model, and the overall inter-annotator agreement for the annotations was 0.83 Cohen's Kappa ($\kappa$), indicating excellent agreement. I describe in detail the annotation scheme and procedure in Chapter §5, and also release the corpus for use by other researchers.

### 1.4.3 Emotion Lexicon

To have an appropriate emotion lexicon consistent with the Plutchik's 8-emotion psychological model [Plutchik, 1980], I analyzed and improved one of the most commonly used GPELs, namely, the NRC lexicon [National Research Council of Canada; also known as the Emolex emotion lexicon Mohammad et al., 2013, Mohammad and Turney, 2013, 2010]. The NRC used Macquarie's Thesaurus [Bernard, 1986] as the source for terms,

retaining only words that are repeated more than 120,000 times in Google n-gram corpus [Michel et al., 2011]. The NRC maps each word to zero or more labels drawn from Plutchik's model, and provides labels for 14,182 individual words.

While the NRC has been used extensively across the emotion mining literature [Tabak and Evrim, 2016, Abdaoui et al., 2017, Rose et al., 2018, Lee et al., 2019, Ljubešić et al., 2020, Zad et al., 2021d], close inspection reveals a large number of incorrect, non-sensical, pejorative, or otherwise troubling entries. While I provide more examples later in the chapter, to give a flavor of the problem, the NRC provides emotion labels for many generic nouns (*tree*→ANGER), common verbs (*dance*→TRUST), colors (*white*→ANTICIPATION), places (*mosque*→ANGER), relations (*aunt*→TRUST), and adverbs (*scarcely*→SADNESS). Furthermore, the NRC suffers from significant ambiguity because it does not include part of speech categories for the terms: for example, while *console* implies SADNESS in its most common verb sense (as the NRC indicates), in its most common noun sense means a small side table, which probably should have no emotive content. In my analysis, many of these problematic entries seem to stem from a conflation of *emotive* (context-independent) and *affective* (context-dependent) emotion language use: it is as if, during the annotation of Shakespeare's *Macbeth*, the annotators of the NRC marked *hell*→ANGER and *woman*→ANGER because of the bard's highly contextualized statement "Hell hath no fury like a woman scorned": while it is true that this statement is often cited to support an assertion that women are angry people in general, and such a lexicon entry would help in correct marking of the affective implication of this specific statement in this particular context, it does not generalize to all, or even most, uses of the word *woman*. Chapter §4 presents more details of this work.

During my work to revise NRC lexicon and ABBE corpus, I mentored the graduate student, Joshuan Jimenez who helped me in this process.

## 1.5 Outline

The dissertation proceeded as follows. First, I review the literature of emotion detection in natural language processing, (§2). I next presented the systematic evaluation of a framework for unsupervised emotion recognition for narrative text (§3), following which I discussed the Emotion Lexicons and revised NRC emotion lexicon (§4). Next, I reported the data I annotated and guideline (§5). Afterward, I explained the details of the animate beings emotion detection system (§6). I concluded the dissertation by listing my contribution (§7).

Chapter 2

# Related Work

## 2.1 Psychological Emotion Theories

There are many ways of defining emotion, not all of which are relevant to the task of finding emotions in text. For example, we know of theories of emotion that go back to the Ancient Greeks and Romans—such as Aristotle, Cicero, Senaca, and Galen—and emotion remained a topic of theorizing through the Middle Ages (Augustine, Aquinas) and Renaissance (Machiavelli, Montaigne) [Schmitter, 2021]. In the dawn of the scientific age of psychology, thinkers as august as Charles Darwin and William James found emotion to be worthy of their attention and effort [Darwin, 1872, James, 1894].

Modern theories of emotion have three main dimensions of explanation or description: physiological, neurological, and cognitive. According to physiological views, emotions are responses within the human body to external or internal stimuli. According to neuroscientific views, emotional reactions can be explained by neural processes in the brain.

Cognitive approaches, pursued in psychology and cognitive science, have generally been considered the most useful for text processing. The American Psychological Association (APA), for example, defines emotion as *a complicated reaction pattern that can be noted in various ways, where emotion is composed of elements such as behavioral, physiological, and experiential based on how an individual deals with an event that has significance to them* [VandenBos, 2007]. The Dictionary of Cognitive Psychology [Eysenck et al., 1994], on the other hand, does not formally define emotion, but an operative definition emerges from its five pages devoted to emotion: emotion is a *mental state*. Cognitive theories of emotion vary in their complexity, with some theories identifying sophisticated constellations of components, including the activation of appraisals, the holding of subsequent desires, and the formation of intentions [Izard, 1992]. For some

theorists, all cognition participates more or less in emotion [Scherer, 1993]. Despite this range of complexity, what is critical for this work is that emotion is a mental state that must be attributed to a being capable of maintaining such a state.

Also, modern psychological theories of emotion may be grouped into two types: *categorical* and *dimensional* [Calvo and Mac Kim, 2013]. However, based on my observations, psychological theories of emotion can be divided into three types: *categorical* or *discrete*, *dimensional*, and *hybrid*.

Categorical or discrete psychological models represent basic emotions as individual, distinct categories, e.g., Oatley and Johnson-Laird's (1987) with five basic emotions, several models with six basic emotions [Ekman, 1992a, Shaver et al., 1987], Parrott's model of six basic emotions arranged in a three-level tree (2001), Panksepp's model with seven emotions (1998), and Izard's with ten (2007).

Dimensional psychological models, by contrast, determine emotions by locating them in a space of dimensions (usually two to four) that might include arousal, valence, intensity, etc. These include two dimensional models such as Russell's circumplex model (1980), Scherer's augmented circumplex (2005), and Whissell's model [Whissell, 1989]. Lövheim's model (2012) is an example that uses three dimensions, while Ortony et al. [1990], Fontaine et al. [2007], Cambria et al. [2012] proposed four-dimensional models.

Finally, there are also models which combine both categorical and dimensional aspects; these are *hybrid* models, the most prominent of which is Plutchik's wheel and cone model with eight basic emotions [Plutchik, 1980, 1984, 2001a].

Of all the many emotion models that have been proposed, Ekman's 6 category model (anger, disgust, fear, happiness, sadness, and surprise) is by far the most popular in computational approaches, partly because of its simplicity, and partly because it has been successfully applied to automatic facial emotion recognition [Zhang et al., 2018, Suttles and Ide, 2013, Ekman, 1992a,b, 1993]. This is despite that some researchers have doubts

that Ekman's model is complete, as it seems to embed a Western cultural bias [Langroudi et al., 2018]. In my own review of emotion recognition systems, as discussed in (§3), the highest performing system reported for narrative text was described by Kim et al. [2010]. In that work, they used a four-label subset of Ekman's model (happiness, anger, fear, and sadness), and this is the model I adopted in (§3). In this review, I'll go through psychological theories in significant detail.

### 2.1.1 Ekman

Previously Ekman defined a model of emotions, and described the basic constituents as being: anger, disgust, fear, happiness, sadness, and surprise. Ekman's model is very useful for the application of facial emotion detection. However, there is doubt that Ekman's model completely represents the full spectrum of emotions, and that it is limited in that it solely considers the emotions of Western cultures [Langroudi et al., 2018].

### 2.1.2 Parrot

Parrot's model, as shown in Fig. 2.1, also considers six basic emotions, consisting of fear, sadness, surprise, anger, love, and joy [Parrott, 2001]. However, Parrot extended these larger categorizations and arranged them in a tree, that ultimately encompasses 100 separate emotions.

### 2.1.3 A Circumplex Model

As a set of independent and correlated affective dimensions, Russell introduced a circumplex model as shown in Fig. 2.2a in two dimensions along two orthogonal axes which plots 150 affective labels. Arousal (activation) in the vertical axis represents deactivation

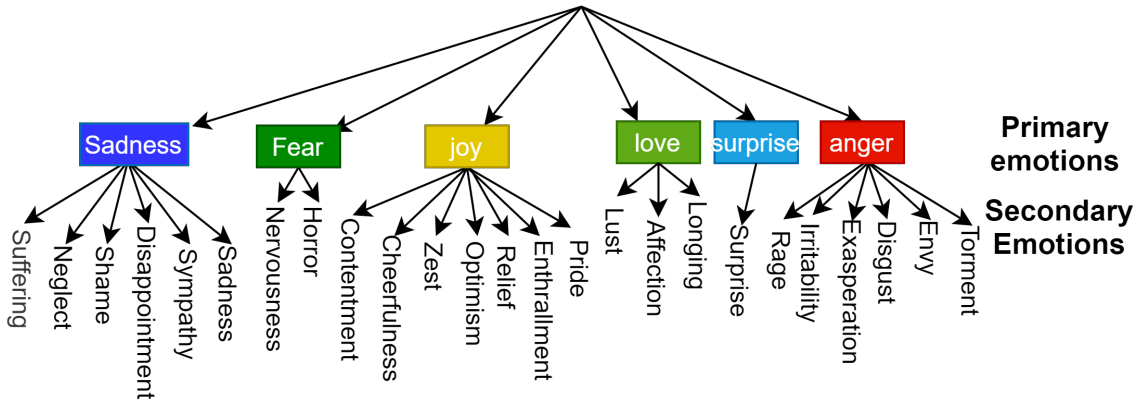
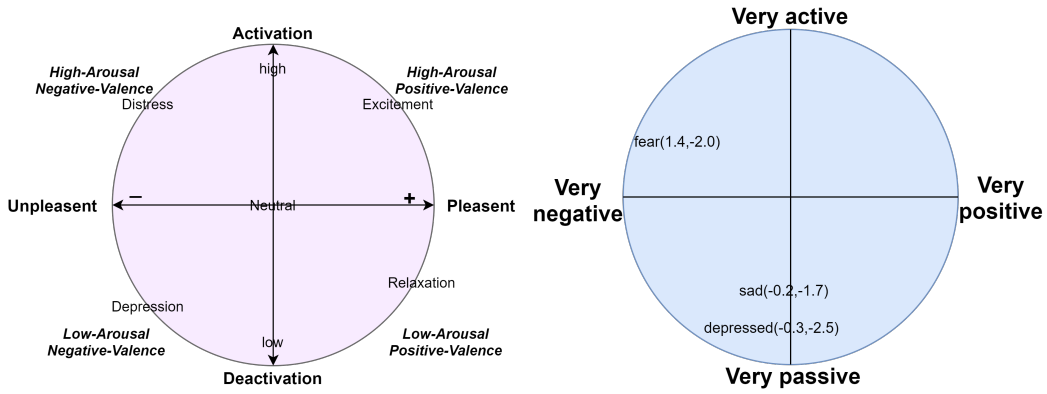



Figure 2.1: Parrot's emotions model

in negative, neutral in center, and activation in the positive side, whereas the horizontal axis performs unpleasant in the negative side, medium in center, and pleasant in the positive side. This circumplex model includes pleasure, excitement, arousal, distress, displeasure, depression, sleepiness, and relaxation[**?**Russell, 1980]. Russell et al. described the affective dimensions (anger, fear, shame, jealousy, etc.) as horizontally as part of circumplex and vertically as a fuzzy hierarchy. They believed emotion can be divided into comprehensible entities [Russell and Barrett, 1999]. In this model, similar emotions like frustrated, distressed, and annoyed are grouped close together and dissimilar emotions are placed further apart. Unlike Ekman's model, Russell's model points to relationships between emotions. For example, with feeling depressed there is an expectation to experience little to no feelings of happiness [Langroudi et al., 2018].

### 2.1.4 Scherer's Update to the Russell's Model

Russell's model was used to classify emotions exactly on an edge of circumplex shape, which for every x and y coordinates applies the equation of $X^2 + Y^2 = R^2$. However, it does not cover the emotions in the circle. To remedy this concern, Scherer created a model that for every x and y emotions there exists equation $(x_h)^2 + (y_k)^2 = R^2$, which




(a) Circumplex's emotions model (b) Whissell's emotions model

Figure 2.2: Examples of 2D psychological emotion models, Circumplex (a) and Whissell models (b)

covers the whole circle. In this way, the model covers the emotions, for example, zero valence and small negative level of arousal. This model is bi-dimensional of arousal and valence and illustrates a broad range of emotions [Langroudi et al., 2018, Scherer, 2005]. Georgies et al. also used Sherer's model for classifying blog posts in two dimensions of emotion analysis to predict the level of valence and arousal of each text [Perikos and Hatzilygeroudis, 2013].

### 2.1.5 Whissell

Whissell suggested a continuous bi-dimensional emotion model as illustrated in Fig. 2.2b whose dimensions are evaluation and activation in a pair of <activation, evaluation> which are assigned to the words from the Dictionary of Affect (approximately 9000 words). The evaluation dimension is about pinpointing feeling from negative to positive, and the activation dimension is measured by taking action from active to passive [Whissell, 1989].

### 2.1.6 Plutchik

Plutchik proposed a three-dimensional hybrid model as shown in Fig. 2.3 which consists of eight basic-complex emotions (joy, sadness, anger, fear, trust, disgust, surprise, and anticipation). He arranged emotions like a color wheel in four groups of primary, secondary, tertiary, and opposite emotions. In the emotion wheel, basic emotions are located in the inner portion and become more complex in the outer portions. Each of them is further sub-divided into three ranges of emotions, where those close to the inner circle are more intense. For example, anticipation is more complex than vigilance but less intense. Interest is more complex than anticipation and less intense. Also, similar emotions are organized closely together, and opposite emotions are 180 degrees apart from each other [Plutchik, 2001a]. Among different psychological emotion models, Plutchik's wheel model has received the most attention within natural language processing, and this is the model I choose for (§4), (§5), and (§6) .

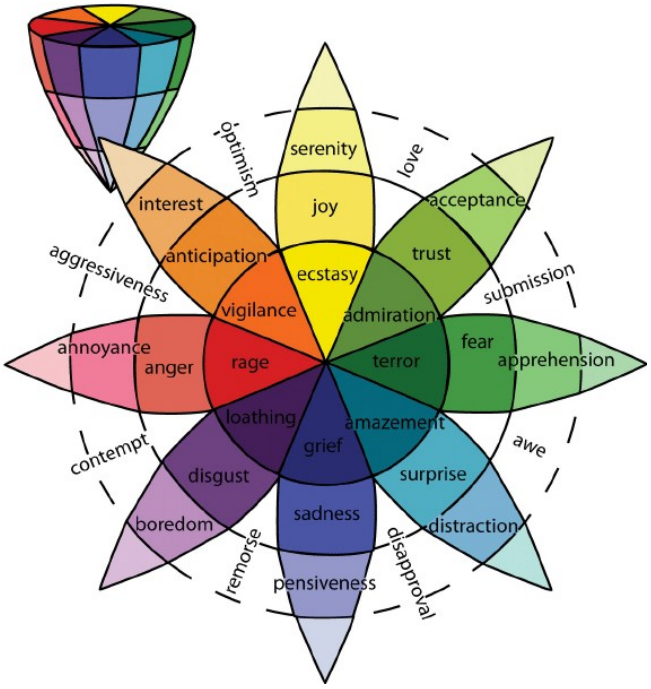


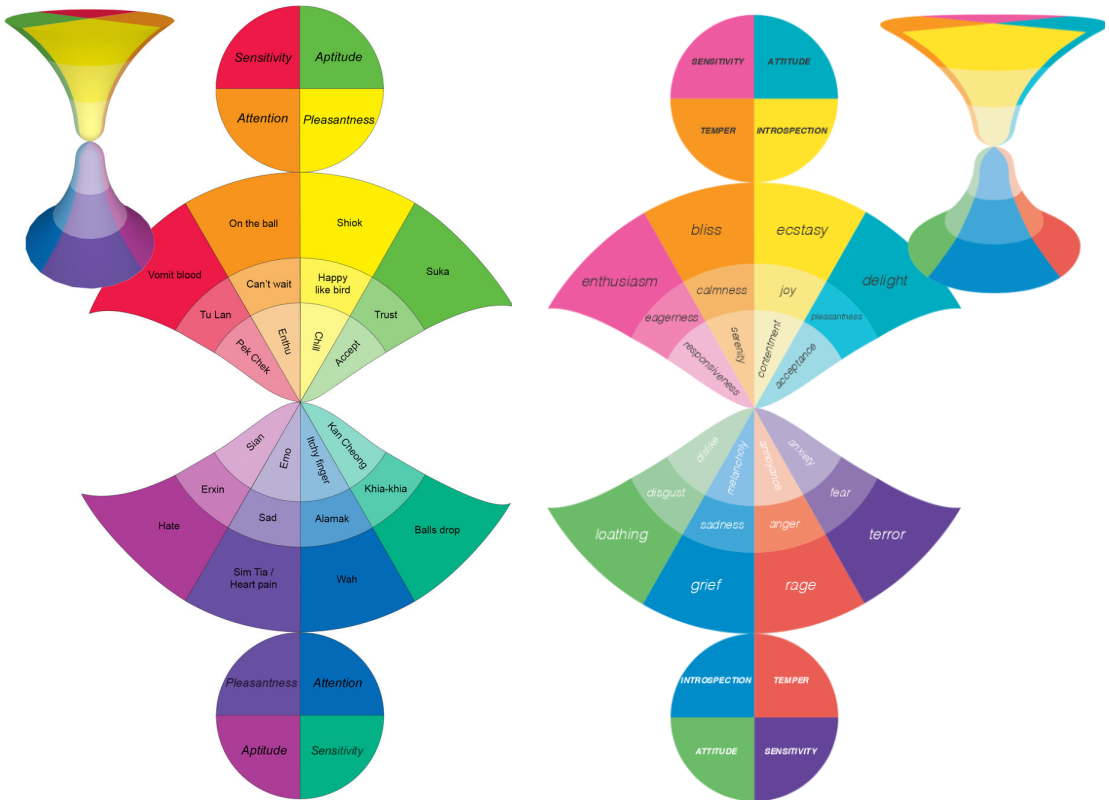

Figure 2.3: Plutchik's emotions wheel, [Plutchik and Conte, 1997]. Figure taken from [Maupome and Isyutina, 2013], with permission.

### 2.1.7 OCC Model

Ortony et al. offered an OCC model consisting of 22 emotion categories in three sources: goals, standards, and tastes, and each of them are the basis of three types of events, actions, and objects. Hence, events describe desirability or undesirability of goals, actions explain praise or blameworthiness of standards, and objects reflect appealing or unappealing tastes of individuals. Moreover, some emotions are caused by mixing two or three types of emotions [Ortony et al., 1990, Clore and Ortony, 2013]. Steunebrink et al. critiqued the OCC model, and offered a computer scientist's perspective, suggesting ways to make it more useful, practical, and computable. It was identified some ambiguities and proposed a new structure of emotions which is more practical in AI [Steunebrink et al., 2009]. Moreover, It is considered as a standard composite emotion model [Perikos and Hatzilygeroudis, 2016].

### 2.1.8 Hourglass of Emotions

The Hourglass model of emotions is shown in Fig. 2.4a, and consists of 20 categories (half positive and half negative) in four independent dimensions [Cambria et al., 2012]. It is based off of Plutchik's model. Each emotion is presented in a pair of words in order to indicate the root of the word. Cambria et al. claimed that their proposed model can explain the entire emotional experience that happens to everyone. As an example, the difference between guilt and shame is the negativity between the self and an act. Guilt is caused by believing a bad thing was done. Shame is caused when an individual thinks they are a bad person. The difference is very important because it displays distinct results and satisfactorily represents similarities and differences of the effective words. In 2020 based on some empirical findings in the context of sentiment analysis, **?** revisited the Hourglass of Emotions, an emotion categorization model optimized for polarity detection.




(a) Hourglass's emotions model, [Cambria et al., 2012]. Figure taken from [Cambria et al., 2012].

(b) Revisited Hourglass's emotions model, [Susanto et al., 2020]. Figure taken from [Susanto et al., 2020].)

Figure 2.4: Examples of 4D emotion models, Hourglass (a) and Revisited Hourglass (b)

### 2.1.9 Fontaine

Fontaine et al. studied three languages and found that the four-dimension emotional model can present the full range of emotional experiences in everyone. The four spaces that they applied on the three languages are evaluation- pleasantness, potency- control, activation- arousal, and unpredictability. They studied six major emotion components that, in total, comprise 144 features [Fontaine et al., 2007].

## 2.2 Emotion Lexicons

A psychological theory of emotion usually goes hand-in-hand with an emotion lexicons and one of the key language resources for emotion detection in text is an emotion lexicon, which is simply a list of words associated with emotion categories. Emotion lexicons take a specific emotional theory and associate the labels or values in that theory with specific lexical entries. Emotion lexicons can be used both in rule-based and machine-learning-based approaches to emotion detection. There are two types of emotion lexicons. One is general purpose emotion lexicons (GPELs) which specify the generic sense of emotional words. GPELs sometimes express emotions as a score, and can be applied to any domains. Prominent GPELs include WordNet Affect [WNA; Strapparava and Valitutti, 2004], the Wisconsin Perceptual Attribute Rating Database [WPARD; Medler et al., 2005], Linguistic Inquiry and Word Count [LIWC; Pennebaker et al., 2001], and the National Research Council (NRC) and NRC Hashtag lexicons [Mohammad and Turney, 2010, Mohammad et al., 2013]. The second type of lexicon are domain specific emotion lexicons (DSELs) which are targeted at specific domains for emotion recognition. Bandhakavi et al. [2014], for example, proposes a domain-specific lexicon for emotional tweets. Table 2.1 compares the details of several key GPELs.

There are lexicons which are related to emotion, but not themselves emotion lexicons. For example, Staiano and Guerini [2014] described the DepecheMood lexicon, which was an automatically generated, general-purpose, and mood lexicon with 37K terms. It includes eight mood-related labels (*don't care*, *amused*, *annoyed*, *inspired*, *anger*, *sadness*, *fear*, and *joy*) based on Rappler's mood meter (obtained by crawling the `rappler.com` social news network). Kušen et al. [2017] compared the four labels shared between NRC and DepecheMood (anger, sadness, fear, and joy), and showed that NRC had the highest recall. NRC performed better at capturing fear, anger, and joy, and

| Emotion Lexicons | Citation | Set of Emotions | Entries |
|---|---|---|---|
| Revised NRC | Zad et al. [2021c] | Plutchik basic model 1980 | 6,166 (wordforms+POS) |
| NRC Hashtag | Mohammad et al. [2013] | Plutchik's basic model | 32,400 |
| NRC / Emolex | Mohammad and Turney [2010] | Plutchik basic model 1980, neg./pos. | 14,182 |
| WPARD | Medler et al. [2005] | positive or negative | 1,402 |
| WNA | Strapparava and Valitutti [2004] | a hierarchy of emotions | 915 synsets |
| LIWC | Pennebaker et al. [2001] | affective or not, neg./pos. anxiety, anger, sadness | 5,690 |
| ANEW | Bradley and Lang [1999] | 3D (valence, arousal,dominance) | 1,035 |
| General Enquirer | Stone et al. [1966] | pleasure, arousal, feeling, pain | 11,788 |

Table 2.1: Emotion-related lexicons table. WNA= WordNet Affect; NRC= National Research Council in Canada; LIWC= Linguistic Inquiry and Word Count; WPARD= Wisconsin Perceptual Attribute Rating Database; ANEW= Affective Norms of English Words

DepecheMood performed better at recognize sadness. Araque et al. [2019] created the extended DepecheMood++ (DM++) for English on Rappler news and Italian on Corriere news (`corriere.it`, an online Italian newspaper).

Table 2.1 lists the main emotion lexicons in details. As can be seen, the NRC is one of the largest resources and uses one of the more expressive emotion ontologies, hence researchers' preference for it in their work.

### 2.2.1 NRC & Revised NRC Emotion Lexicons

For Plutchik's model, the most prominent general purpose emotion lexicon is the *NRC* (National Research Council of Canada) emotion lexicon also known as *Emolex*. It is a "word-sense level" emotion lexicon and comprises 14,182 words labeled according to Plutchik's psychological model [Plutchik, 1980]. It is a General Purpose Emotion Lexicon (GPEL) derived from widely available sources and applicable to all domains [Zad et al., 2021b], and manually annotated through Amazon's Mechanical Turk service into Plutchik's eight basic emotions. The NRC was created via a crowd-sourcing, and used Roget's Thesaurus as the source for terms [Mohammad and Turney, 2010, 2013, Mohammad et al., 2013]. Because Chapter (§4) focuses on NRC terminology, I go over it in depth there and explain that I discovered a substantial number of biased entries, prompting me

to revise the NRC emotion lexicon and develop the Revised NRC emotion lexicon. The result is that Zad et al. [2021c] revised the NRC emotion lexicon semi-automatically for correcting problems such as disambiguating POS categories, a large number of troubling, inaccurate, nonsensical and pejorative entries, and it is this *Revised NRC* that I use in (§6).

### 2.2.2 NRC Hashtag Emotion Lexicon

NRC Hashtag Emotion Lexicon [Mohammad, 2012] comprises 16,862 words, drawn from Twitter hashtags, that are labeled with a strength of association (from 0 to infinity) for each of six emotion classes. It was created automatically by extracting tweets that contains `#joy`, `#sadness`, `#surprise`, `#disgust`, `#fear`, and `#anger`. Mohammad [2012] showed that the NRC Hashtag emotion lexicon provides better performance on Twitter Emotion Corpus than the WordNet-Affect emotion lexicon, but not as good as the original NRC emotion lexicon. Mohammad and Kiritchenko [2015] extended this work by expanding the hashtag word list to 585 emotion words, producing 15,825 labeled entries, with performance on headline data set again better than WNA.

### 2.2.3 WordNet Affect Version 1.1

The WordNet Affect Lexicon [WNA or WAL Strapparava and Valitutti, 2004] is an emotion lexicon based on WordNet [Fellbaum, 1998b]. WNA classifies 289 WordNet *Noun* synsets (a group of synonym words that express a notion) into an emotion hierarchy rooted in an augmented version of Ekman's basic emotions, and partially depicted in Figure 2.5. WordNet links an additional 1,191 *Verb*, *Adverb*, and *Adjective* synsets to this core *Noun*-focused hierarchy. These synsets represent approximately 3,500 English lemma-POS pairs. Kim et al. used WordNet Affect, which builds upon the general WordNet database [Fellbaum, 1998a], and I refer to it extensively in (§3).

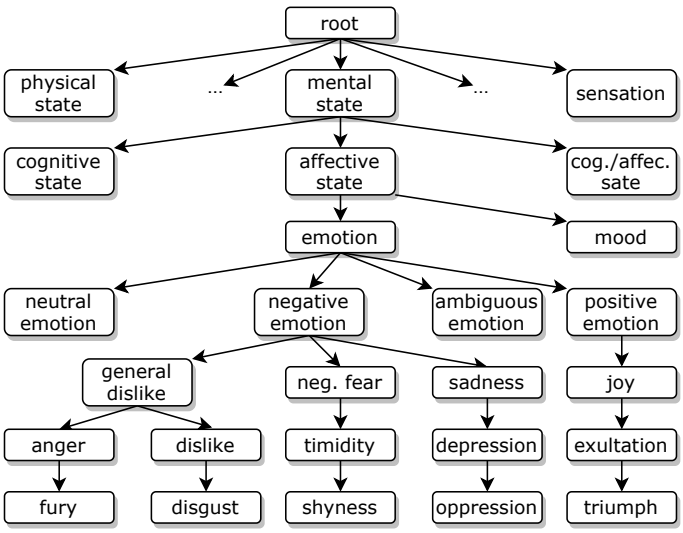



Figure 2.5: Hierarchy of emotions in WordNet Affect Version 1.1.

### 2.2.4 General Enquirer Emotion Lexicon

The General Enquirer lexicon, while not specifically designed as an emotion lexicon, comprises 11,788 concepts labeled with 182 category labels that includes certain affect categories (e.g., pleasure, arousal, feeling, and pain) in addition to positive/negative semantic orientation for concepts [Stone et al., 1966].

### 2.2.5 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count [LIWC Pennebaker et al., 2001, 2007] is a text analysis program that includes a lexicon comprising 2,300 entries spread across 73 categories, many of which are emotive or have sentiment, including NEGATION, ANGER, ANXIETY, SADNESS, etc.

## 2.3 Emotion Datasets

Annotated corpora of emotion-laden language go hand-in-hand with emotion lexicons. This is because one of the first tests of the utility of a lexicon is how well a system that uses the lexicon performs on automatic labeling. In general, data annotation is a crucial part of most machine learning research and affects the quality of the work substantially. As is commonly known, in the case of linguistic annotation, manually labeling large amounts of text is expensive and time consuming; further, in most cases, assigning labels can be subjective and dependent on the personality, emotions, background, and point of view of the annotator; and finally, unbalanced label frequency creates challenges for training various learning algorithms.

There are several text corpora annotated with emotional categorical models [Yadollahi et al., 2017, Sailunaz et al., 2018, Acheampong et al., 2020]. For example, the International Survey on Emotion Antecedents and Reactions (ISEAR) corpus Scherer and Wallbott [1994] comprises 7,665 sentences drawn from 3,000 students from 37 countries were asked to report as a sentence or paragraph situations in which they had experienced FEAR, SADNESS, JOY, ANGER, SHAME, GUILT, and DISGUST emotions. ISEAR data set is annotated by authors and labeled by seven emotions (FEAR, SADNESS, JOY, ANGER, SHAME, GUILT, and DISGUST). Similarly, Aman's corpus Aman and Szpakowicz [2007] comprises of 1,466 sentences from blogs and labeled by seven emotions (SADNESS, SURPRISE, ANGER, FEAR, DISGUST, HAPPINESS, and MIXED EMOTIONS). The Semantic Evaluations (SemEval) corpus [Rosenthal et al., 2019] includes 1,250 news headlines and labeled by Ekman's six basic emotions (ANGER, DISGUST, SURPRISE, FEAR, JOY, and SADNESS). These are just three examples of many. Table 2.2 presents most of emotion data sets in detail.

| Name | Year | Author | Size | Type of Data |
|---|---|---|---|---|
| ISEAR | 1997 | Scherer | 7,666 | Sentences |
| Fairy tales | 2005 | Alm | 15,302 | Sentences |
| SemEval | 2007 | Strapparava | 1,250 | News headlines |
| TEC | 2012 | Mohammad | 21,000 | Tweets |
| CBET | 2015 | Shahraki | 81,163 | Tweets |
| EmoBank | 2017 | Buechel | 10,548 | Sentences |
| CrowdFlower | 2016 | crowdsourcing | 39,740 | Tweets |
| Blogs | 2007 | Aman | 5,205 | Sentences |
| DailyDialogs | 2017 | Li | 13,118 | Sentences |
| Electoral-Tweets | 2015 | Mohammad | 100,000 | Tweets |
| EmoInt | 2017 | Mohammad | 7,097 | Tweets |
| Emotion-stimulus | 2015 | Ghazi | 820 | Sentences |
| FB-valence-arousal | 2016 | Preot¸iuc-Pietro | 2,895 | Facebook posts |
| Grounded-Emotions | 2017 | Liu | 2,557 | Instances |
| SSEC | 2017 | Schuff | 4,868 | Tweets |
| Project Gutenberg | 2009 | Lebert | 34,000 | Books |
| Google | 2011a | Michel | 5.2 million | Digitized books |
| Hashtag Emotion Corpus | 2015 | Mohammad | 21,051 | Tweets |

Table 2.2: Emotion-related data sets

### 2.3.1 Alm's Fairy Tales

Alm's fairy tale corpus conducted in 2005 by Cecilia Alm contains 15,302 sentences from 176 children's fairy tales from classic collections by Beatrix Potter, the Brother's Grimm's, and Hans C. Andersen. Two annotators marked both the emotion and mood of each sentence in the corpus (i.e., two separate judgements by both annotators, for a total of four labels per sentence), using Ekman's six emotions (JOY, FEAR, SADNESS, SURPRISE, ANGER, and DISGUST). 1,167 sentences in the corpus had "high annotation agreement" which Alm defined as all four labels being the same, and there are a total of 4,627 other sentences which annotators have all labeled them as neutral. One reason to focus on only the high agreement sentences is because the overall Cohen's Kappa for the dataset agreement is a quite poor -0.2086. If we focus only on high agreement, the Cohen's Kappa is perfect [Alm et al., 2005, Alm, 2010]. From this, Alm was able to work on a total of 1,580 sentences with emotion labels that allowed her to help detect emotion

in novels using supervised machine learning with the SNoW learning architecture. Emotion annotation is notoriously difficult, and very few emotion annotation projects have achieved high agreement. This suggests that most of the approaches to emotion annotation have suffered from lack of conceptual clarity. I chose this corpus in (§4) because of the ready availability of an emotion detection system [Zad and Finlayson, 2020] that uses this corpus for evaluation.

### 2.3.2 ISEAR

The International Survey on Emotion Antecedents and Reactions dataset [Scherer and Wallbott, 1994] is a collection of student responses when they were asked to report situations that occurred to them in which they had experienced 7 major emotions (Joy, Fear, Anger, Sadness, Disgust, Shame, and Guilt). Approximately 3,000 respondents in 37 countries responded detailing the way they had appraised the situation and their reaction towards it.

### 2.3.3 SemEval

This workshop is dedicated to performing semantic analysis on text regarding various topics [Zad et al., 2021e]. An important topic to focus on is sentiment analysis, where text from outlets like Twitter [Rosenthal et al., 2017] or other social media platforms [Patwa et al., 2020] are parsed to check for whether a certain message is regarded as positive, negative, etc. These topics are done as tasks, which help create high-quality annotated datasets. These datasets can be used to help assist in other systems to help add more semantic information.

### 2.3.4 EmoBank

Buechel and Hahn [2017] created the EmoBank dataset, a text corpus that was manually annotated based on the Valence-Arousal-Dominance scheme. This was done by collecting a large number of blogs, essays, news headlines, and other types of text to create a corpus that has around 10,548 English sentences. The dataset is annotated for Ekman's model of basic emotion (Anger, Sadness, Fear, Joy, Surprise, and Disgust).

### 2.3.5 Emotion-Stimulus

The emotion-stimulus dataset [Ghazi et al., 2015] differs from the other datasets in that it is based on FrameNet's *emotions-directed* frame while also noting what was the cause of the emotion being felt in a sentence. This dataset contains 820 sentences with both annotations while 1,594 sentences only have the emotion tags (Happiness, Sadness, Anger, Fear, Surprise, Disgust, and Shame, which is the Ekman model plus the Shame label).

## 2.4 Emotion Detection Approaches

There is a social science/humanities hypothesis that the "emotionality" of texts was quite different in the period of revolutionary France [Tackett, 2015]. For example, let's assume that we have three stories: the first is written in the 17th century, the second is written in the 19th century and the third one is contemporary, all sharing a word. By applying an emotion detector on each document, we can find three types of labels for that word. By comparing the three labels, we can find out how the emotion of a word has evolved in time. However, automatically identifying emotions expressed in text (a.k.a. text emotion detection) has received a lot of attention recently in the natural language processing world, and it is a relatively new technology.

There have been at least one hundred papers describing approaches to emotion recognition in text [Calefato et al., 2017, Teng et al., 2007, Shaheen et al., 2014]. Text-Based Emotion Detection (TBED) has been used by researchers to automatically detect affect, identifying the feelings and sentiments expressed in a text. It detects emotions from a variety of data sources using computational linguistics, text analysis, machine learning, and natural language processing (NLP). These systems can be roughly separated by the usual distinction between statistical and rule-based approaches, with a few hybrid systems being available. Some use general learning and statistical approaches to find valuable features based on the theoretical rules and experimental evidence existing in the corpus (rule-based); while others would prefer using lexical and semantic analysis to extract features by better understanding of semantic and grammatical rules and features, which they should rely on the output of humans and sometimes linguistics who know the language pretty well (statistic-based). Finally, there are tasks that use a hybrid of the two approaches based on data availability, accuracy and precision. Here I review a selection of approaches that have been applied to narrative-like or narrative-related discourse types. It is important to remember that all of these approaches use different data and different theories, often involving different numbers of labels. All things being equal, classification results usually degrade as the number labels increases; therefore the performance of each system can only be loosely compared.

Kozareva et al. [2007] studied a headline emotion classification from the World Wide Web based on frequency of words collected from MyWay, AlltheWeb, and Yahoo. They claimed words that have a high frequency through many texts with a given emotion have a possibility to express the emotion. They used six emotions for this study which are angry, fear, sadness, surprise, disgust, and joy. They combined all the frequency word counts of the three resources and measured MIs (Mutual Information Scores) of a bag of content and emotion words. Their proposed model can verify the predominance of a sentiment.

The results show that most of the correct sentiment assignments are related to negative emotions. The highest accuracy for disgust is 97.3 and the lowest accuracy is 75.30 for fear.

Strapparava and Mihalcea [2008] described a system for recognizing emotions in news headlines. They extracted 1,250 news headlines from a variety of news websites (such as Google news, CNN, and online newspapers) and annotated them using Ekman's model—JOY, FEAR, SADNESS, SURPRISE, ANGER, and DISGUST—splitting the data into a training set of 250 and a test set of 1,000 (this is called the *SemEval-2007* dataset). They tested five approaches: WNA-PRESENCE, LSA-SINGLE-WORD, LSA-EMOTION-SYNSET, LSA-ALL-EMOTION-WORDS, and NAIVEBAYES-TRAINED-ON-BLOGS. WNA-PRESENCE, which looked for headline words listed in WNA, provided the best precision at 0.38. The LSA-ALL-EMOTION-WORDS, which calculated the vector similarity between the six affect words and the LSA representation of the headline, led to the highest recall and $F_1$, at 0.90 and 0.176, respectively.

Aman and Szpakowicz [2008] used a Support Vector Machine (SVM) trained and tested on blog data for recognition Ekman's emotion classes (JOY, FEAR, SADNESS, SURPRISE, ANGER, and DISGUST), plus two additional classes: *mixed emotion*, and *no emotion*. Four human judges manually annotated 1,890 sentences from automatically retrieved blogs to create the corpus. The features for the SVM were the presence of emotion words listed in Roget's thesaurus and WNA. $F_1$ measures for each emotion class ranged between 0.493 to 0.751, in each case surpass the baseline performance.

Tokuhisa et al. [2008] described a lexicon-based emotion recognition system for Japanese. They handcrafted emotion lexicon by identifying 349 emotion words from the Japanese Expression Evaluation (JEE) Dictionary classified into 10 different emotions: 3 positive (HAPPINESS, PLEASANTNESS, RELIEF) and 7 negative (FEAR, SADNESS, DISAPPOINTMENT, UNPLEASANTNESS, LONELINESS, ANXIETY, and ANGER). They then used

this lexicon to automatically assemble a labeled corpus of 1.3M emotion-provoking (EP) "events" (defined as a subordinate clauses which modifies an emotional statement). They then demonstrated a two-step method for emotion recognition, starting with SVM-based coarse sentiment polarity classification (positive, negative, or neutral) followed by kNN-based classification of non-neutral instances into the appropriate fine-grained emotion classes (3 for positive, 7 for negative). Their reported accuracies of between 0.5 and 0.8 for their best performing model.

Kim et al. [2010] reported the highest performing emotion recognition system on narrative text. Among their data was a set of 176 fairy tales whose 15,087 sentences were labeled by Alm [2008] with a four-emotion subset of Ekman's theory (anger, fear, joy, and sadness). They demonstrated an unsupervised approach, where each sentence is transformed into a vector in a space of emotion words (drawn from WNA and ANEW), and then compressed using a dimension reduction technique (NMF, LSA, or pLSA). These vectors were then compared to reference vectors in the same space that were computed for each of the four emotions. They reported a performance of $F_1$ of 0.733 for NMF, which was their highest performing model. One advantage of this approach was that it is unsupervised, which means both that significant amounts of training data are not required and that all the annotated data can be used for testing. This is important because of the small size of the corpus on which the technique was tested.

Cherry et al. [2012] presented two supervised machine learning models for emotion recognition in suicide note sentences. They used the 2011 i2b2 NLP Challenge Task 2, which comprised 4,241 sentences (600 documents) in the training set, and 1,883 sentences (300 documents) in the test set, which were manually annotated with 15 emotion labels. They used fifteen emotions (ABUSE, ANGER, BLAME, FEAR, FORGIVENESS, GUILT, HAPPINESS, PEACEFULNESS, HOPEFULNESS, HOPELESSNESS, INFORMATION, LOVE, PRIDE, SORROW, and THANK-FULNESS) and Roget's thesaurus to use synonyms

of those emotions. A one-classifier-per-emotion approach yielded an $F_1$ of 0.55, while a latent sequence model that applied multiple emotion labels per sentence achieved an $F_1$ of 0.53. Their latent sequence model is a multi-label sentence classifier which annotates with zero or more emotions. Their system uses one classifier per emotion, and it simplifies label balance and fast development issues. It is a binary classifier and produces a stronger result. Also, they noted that more than 73% of their training data lacked labels which limited the effectiveness of the training.

A supervised-learning-based emotion detection system that uses an emotion lexicon was proposed by Wang et al. [2012], and this system automatically generates emotion-labeled data sets from Twitter containing about 2.5 million tweets for seven emotions. They applied LIBLINEAR and Multinomial Naive Bayes machine learning classifiers with a 7-class emotion scheme (JOY, SADNESS, ANGER, FEAR, SURPRISE, LOVE, and THANKFULNESS). They used the LIWC dictionary and MPQA lexicon for polarities analysis and WordNet-Affect emotion lexicon with the feature combination of unigrams, bigrams, existing sentiment, and part of speech. The system achieved an accuracy 65.57%.

Another prototypical emotion detection system that uses the NRC [Mohammad and Turney, 2013] specifically is presented by Kim et al. [2018]. Their model comprised an attention-based module and multiple independent Convolutional Neural Networks (CNNs), using the NRC emotion lexicon is used for word-level labeling. They applied the system on SemEval-2018 tweet data [Mohammad et al., 2018] with multiple labels from eleven possible emotions: (ANGRY, ANTICIPATION, DISGUST, FEAR, JOY, LOVE, OPTIMISM, PESSIMISM, SADNESS, SURPRISE, and TRUST). The best result was 59.79% accuracy for English data.

Bandhakavi et al. [2017] experimented with unigram mixture models (UMMs) for recognizing emotions in tweets, incident reports, news headlines, and blogs. Each corpus was manually annotated with different emotion theories: 280,000 tweets with Parrott's six

primary emotions [Parrott, 2001], 1,250 news headlines and 5,500 blogs with Ekman's six emotion set, 7000 incident reports from the ISEAR dataset[1] labeled with a seven emotion set. One goal of the study was to compare the utility of domain-specific emotion lexicons with general purpose emotion lexicons (DSELs vs GPELs). They found that combining DSEL lexicon words with n-grams, part of speech tags, and additional words from sentiment lexicons yielded the highest performance of 0.60 $F_1$ on the blog data.

Finally, a recent emotion detection system that reports state of the art results using the WordNet Affect [WNA; Strapparava and Valitutti, 2004] emotion lexicon is represented by Zad and Finlayson [2020]. That system applied unsupervised emotion detection techniques on 176 Alm's fairy tales 2008 for four emotions (JOY, SADNESS, FEAR, and ANGER). The paper explored several different classification techniques, with Non-negative Matrix Factorization (NMF) performing the best with an overall 80.9 $F_1$ score.

Zad et al. applied Non-negative Matrix Factorization (NMF), Principle Component Analysis (PCA), and Latent Dirchelet Allocation (LDA) on Alm's Fairy tales data set, and respectively the overall $F_1$ are reported 80.9%, 76%, and 60.1%. That system applied unsupervised emotion detection techniques on 176 Alm's fairy tales 2008 for four emotions (JOY, SADNESS, FEAR, and ANGER). They applied dimension reduction methods on 1,090 sentences of Alm's fairy tales that all agreed on labels by two annotators from two categories of emotion and mood. They applied WordNet Affect [WNA; Strapparava and Valitutti, 2004] emotion lexicon and augmented it for four emotions [Zad and Finlayson, 2020].

One of the many applications of learning methods in NLP [Zad et al., 2021a, Hajibabaee et al., 2021, Malekzadeh et al., 2021, Heidari et al., 2021a,b,c] is emotion detection. Chiorrini et al. applied deep learning and evaluated Bidirectional Encoder Representations from Transformers (BERT) model on Twitter data, and reported 89% $F_1$ for

[1] `http://www.affective-sciences.org/researchmaterial`

| Citation | Corpus | Lexicon | # Emotions | Method | $F_1$ |
|---|---|---|---|---|---|
| Zad and Finlayson [2020] | Alm's Fairy tales | WNA | 4 | NMF | 80.6 |
| Bandhakavi et al. [2017] | Tweets | UMM+DSEL | 6 | Lexicon only | 0.64 |
| Quan et al. [2015] | Sina | - | 8 | eLDM | 0.64 |
| Sintsova et al. [2013] | Tweets | Olymplex / PMI-Hash | 20 | Dystemo | 0.41 |
| Wang et al. [2012] | Tweets | LIWC and MPQA | 7 | LIBLINEAR and MNB | 0.66 |
| Mohammad [2012] | Tweets | TEC | 6 | ngrams | 0.50 |
| Cherry et al. [2012] | Suicide notes | - | 15 | SVM+LS | 0.55 |
| Kim et al. [2010] | Alm's Fairy tales | WNA | 4 | NMF | 0.73 |
| Aman and Szpakowicz [2008] | Blog | - | 6 | Unigrams | 0.57 |
| Strapparava and Mihalcea [2008] | Headlines | - | 6 | LSA | 0.17 |
| Tokuhisa et al. [2008] | "EP" Events | JEE Dict. | 10 | SVM+kNN | 0.5–0.8 Acc. |

Table 2.3: Emotion recognition approaches on text. LSA = Latent Semantic Analysis; LS = Latent sequence modeling; NMF = Non-negative matrix factorization

emotion detection on four emotions [Chiorrini et al., 2021]. Then Akhtar et al. developed three deep learning models based on CNN, LSTM, and GRU and one supervised model based on SVR. The pre-trained word embedding models of GloVE and word2vec were trained. The proposed model was evaluated on the benchmark setup of EmoInt-2017 and SemEval-2017. They achieved the best accuracy of 74.8% on generic tweets for emotion intensity prediction [Akhtar et al., 2020]. Next, Krommyda et al. applied LSTM machine learning methods and compared them with five other classifiers. It is achieved 91.9% accuracy result for Text-Based Emotion Detection based on Plutchik's emotions in a short text [Krommyda et al., 2021].

## 2.5 Language Resources for Animate Beings

Animacy is the characteristic of independently carrying out actions in the story world [e.g., movement or communication, Jahan et al., 2018a], which includes the ability to experience emotions. Until recently animacy was treated as a word classification task [Bowman and Chopra, 2012a, Wiseman et al., 2015, Lee et al., 2013], which presents problems for identifying referring expressions and coreference chains that refer to animate beings. In contrast, Jahan et al. [2018a] approached the problem as one of marking animacy on coreference chains, which is a more natural fit to the concept of animacy in

stories. In this work, they compiled a corpus using various resources such as Russian Folktales, Islamic Extremist Texts, Islamic Hadiths. In total, the corpus worked on consists of 142 texts, 156,154 tokens, 34,698 referring expressions, and 10,941 coreference chains. They presented a hybrid system merging an SVM classifier and hand-built rules to predict the animacy of referring expressions with an $F_1$ of 0.88, using majority voting to obtain the animacy of coreference chains with an $F_1$ of 0.75. They further extended this work to the detection of characters in narrative [Jahan et al., 2020b], annotating 30 texts from the Corpus of English Novels [De Smet, 2008a] (among other works). I started from these 30 texts to construct the ABBE corpus described (§5).

Bowman and Chopra [2012b] automatically annotated noun phrases based on a taxonomy of ten categories (Human, Org, Animal, Place, Time, etc). The corpus consists of around 600 transcribed dialogues from the parsed part of the Switchboard corpus [Calhoun et al., 2010], which was coded by three undergraduate students from Stanford University. They then leveraged the results of Zaenen et al. [2004] to distinguish which categories were *animate* or *inanimate*. In the process of distinguishing and labeling both inanimate and animate their model achieved an $F_1$ 0.94 (versus a baseline model that labels only animate beings with an accuracy of 0.54).

Chapter 3

# Unsupervised Emotion Recognition for Narrative Text

## 3.1 Introduction

Emotion recognition is a challenging problem on account of the complex relationship between felt emotion and linguistic expression. This includes not only standard natural language processing challenges, such as polysemous words and the difficulty of coreference resolution [Uzuner et al., 2012, Peng et al., 2019], but also emotion-specific challenges such as how context can subtly change emotional interpretations [Cowie et al., 2005]. These technical challenges are exacerbated by a shortage of quality labeled data addressing this task.

Emotion detection tasks are fundamentally predicated on a particular conception of what emotions exist. There has been a significant amount of work on detecting emotion in text [Zad et al., 2021b]. Automatic emotion recognition is useful for many applications. One application of emotion detection includes deriving insights into public opinion of various socio-political topics (e.g. via social media). Another application would be Humanitarian Assistance and Disaster Relief (HADR) efforts that can benefit from emotion detection methods to collect and analyze accurate and reliable information about the disaster situation and people who were affected by the disaster. This is related to the application of emotion extraction on documents or public messages relevant to terrorist attacks to distinguish suspicious activities to better protect our population. Another example application is the automated analysis of historical corpora for better understanding the period in question. The study of product reviews for the purpose of garnering true customer sale prediction and evaluation of products is another potential emotional identification applications.

Among all the applications of emotion detection, my focus is on understanding narratives. A narrative is "a representation of connected events and characters that has an identifiable structure, is bounded in space and time, and contains implicit or explicit messages about the topic being addressed" [Kreuter et al., 2007, p. 222]. By extracting the emotion in a narrative, either general expression of emotion or emotions associated with specific animate beings, we can learn a great deal about the situation being described. For example, the tragedy of *Romeo and Juliet* is a love story by William Shakespeare about two star-crossed lovers; understanding their emotional expressions is critical to understanding the story. Another example is *Titanic*, a great tragic love story, where emotions similarly drive and explain the action. For all these applications, it is crucial to extract emotion from textual data.

These systems can be roughly separated by the usual distinction between statistical and rule-based approaches, with a few hybrid systems being available. However, one constant with these systems is the need to select a psychological theory of emotion as well as rely on an established emotion lexicon of relevant affective terms. With regard to narrative specifically, Kim et al. [2010] reported a high-performing approach to hybrid emotion recognition on a corpus of fairy tales texts [Alm, 2008]. This approach involved an unsupervised learning framework for emotion recognition in textual data, using a modified form of Ekman's psychological theory of emotion [joy, anger, fear, sadness; Ekman, 1992a]. In that work, they used the WordNetAffect (WNA) and ANEW (Affective Norm for English Words) emotion lexicons to construct a semantic space. Each sentence is placed in the space using *tf-idf* weights for emotion words found in the lexicons. They then tested three methods—Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), probablistic Latent Semantic Analysis (pLSA)—for compressing the space to extract features of the constructed vector space model, reduce noise, and eliminate outliers. Finally, the framework used cosine-similarity to label sentences by evaluat-

ing how similar they are compared to standard vectors generated based on WNA entries strongly associated with emotion lexicon (more specifically an extension of WNA). The best performing method was NMF, which they reported achieved an average emotion recognition $F_1$ of 0.733.

Close inspection of the work, however, revealed significant reproducibility problems. Despite my best efforts I were unable to reproduce results anywhere near Kim's reported performance; indeed, my best attempt yielded only roughly 0.25 $F_1$. This was due to several reasons. First, the paper lacked information on model hyper-parameters. Second, the paper omitted descriptions of key NMF steps, including how to identify representative features and what features should be removed before semantic space compression. Third, the paper did not explain how to adapt NMF to deal with the sparse matrices that occur in textual NMF models. Fourth, certain resources associated with WNA either were not correctly identified, or are no longer available. These omissions prevented us from reproducing their models to any degree of accuracy.

Therefore, I undertook to do a systematic exploration of the design space described in Kim et al. [2010]. I examined the highest performing vector space compression techniques reported by Kim *et al.* (NMF), as well as Principle Component Analysis (PCA) and Latent Dirchelet Allocation (LDA) which were reported as high-performing techniques in other work. I show that NMF indeed performs the best, and I clearly explain my experimental setup including methods for identifying relevant features and handling sparse text matrices. The PCA and NMF methods implemented in this chapter are based on the works of Mairal et al. [2009] and Boutsidis and Gallopoulos [2008] respectively which have implemented mechanisms that works for a large sparse matrix (in my case, $1,090 \times 2,405$). This work resulted in an improvement of performance of roughly 7.6 points of $F_1$ over Kim's reported results.

The rest of this chapter is structured as follows. I describe my adapted unsupervised emotion recognition method, giving detailed descriptions of all steps, parameters, and resources needed (§3.2). I next describe the performance of my method on Alm's corpus of fairy tales [Alm, 2008], which was annotated for emotion on a per-sentence level (§3.3). Finally, I identify some unsolved challenges that point toward future work (§6.6), and summarize my contributions (§6.7).

## 3.2 Emotion Recognition Framework

I now describe an unsupervised system for emotion recognition modeled on that reported by Kim et al. [2010]. While I follow the general pattern of that work, I experiment with a different set of dimension reduction methods (NMF from Lee and Seung, as well as PCA and LDA). The system takes as input the following items:

- A corpus containing $n$ sentences $S : s_1, s_2, \ldots, s_n$;
- A set of emotions $E = \{e_1, e_2, \ldots, e_{l-1}, \text{neutral}\}$ for classifying emotions into $l$ different classes, including neutral; and,
- An emotion lexicon $L : \Omega \mapsto E$ which maps each word in the corpus $\omega \in \Omega$ (where $\Omega$ has $m$ terms) to an emotion $e \in E$. The word $\omega$ is in its lemmatized form and has a specific POS. While a lexicon may label a limited number of words, I assume that any words not labeled by the lexicon are implicitly mapped to *neutral*, and thus $\Omega$ is a superset of all the words (in their lemmatized form) present in the given corpus.

A flowchart of the system is shown in Figure 3.1. The system comprises four consecutive steps. In the first step, **pre-processing**, the system processes the input corpus using the CoreNLP library [Manning et al., 2014] to separate the text into sentences and lemmatized tokens. The second step, **vector space modeling**, uses the lemmatized tokens to generate a vector for each sentence in a vector space whose dimensions correspond to

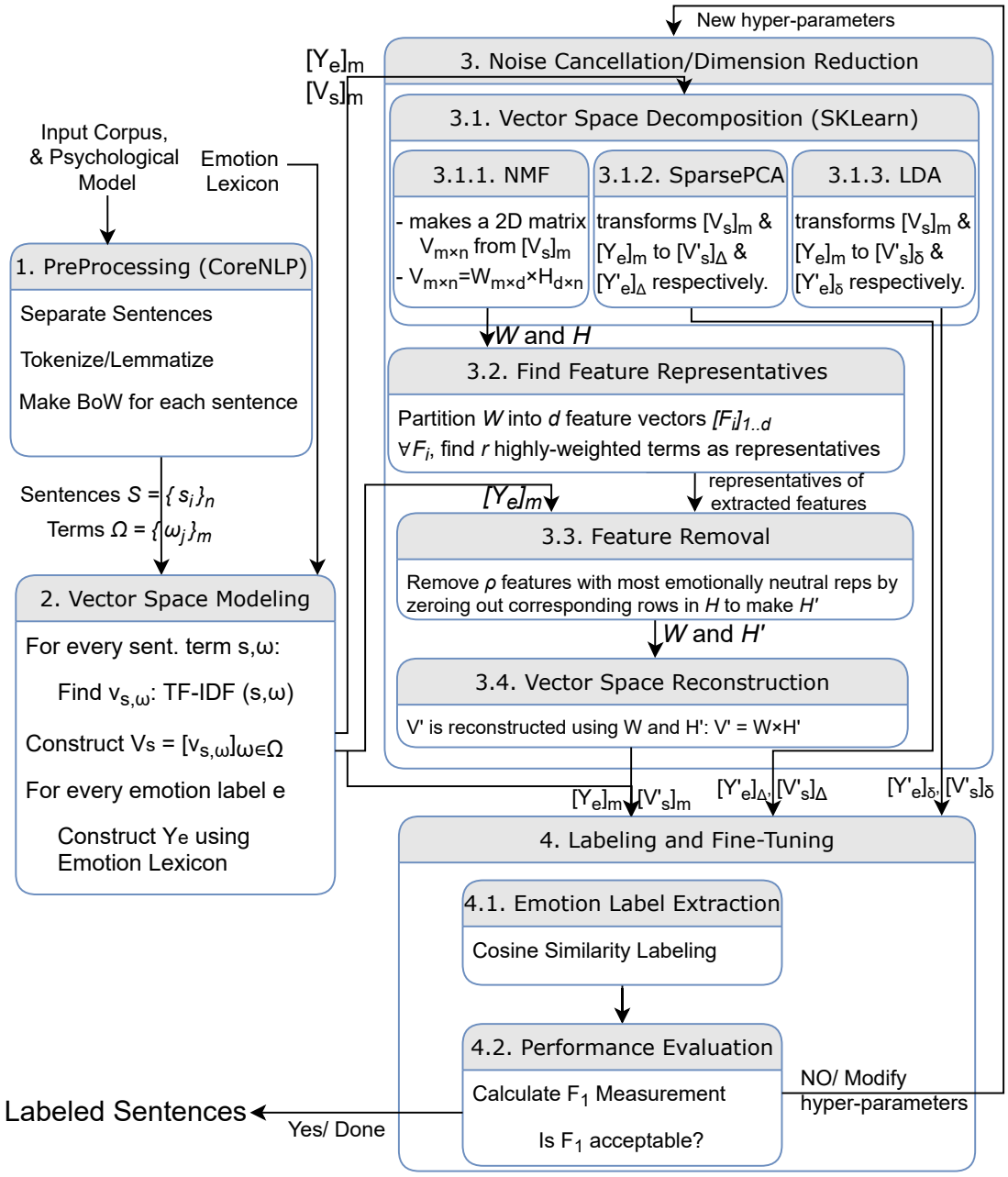



Figure 3.1: Flowchart of the proposed emotion detection system. $[V_s]_m$ and $[Y_e]_m$ represent the original $m$-dimensional sentence and emotion vector model respectively, $[V'_s]_m$, $[V'_s]_\Delta$ and $[V'_s]_\delta$ denote the transformed sentence vector model using NMF, PCA and LDA techniques respectively. $[Y'_e]_\Delta$ and $[Y'_e]_\delta$ denote the transformed emotion vector model using PCA and LDA techniques respectively.

the items in $\Omega$. In the third step, **noise cancellation or dimension reduction**, I explored three different models (Non-negative Matrix Factorization, Latent Dirichlet Allocation, and Principal Component Analysis) to either reduce dimensions or extract features of the vector space. One of my main contributions here is to analyze and explain the effect of this step on the performance of the final emotion recognition system. Finally, the fourth step, **labeling**, compares the vector for each sentence with vectors for each emotion, choosing the closest emotion as the label for the sentence.

**Augmenting WNA** As mentioned before, WNA 1.1 assigns an emotion label to 1,471 synonym sets (synsets) of WordNet. This corresponds to a lexicon of nearly 3,495 affective lemma-POS pairs. Careful inspection of WNA revealed both incorrectly included as well as missing pairs. For incorrectly included pairs, a substantial number were included because all their multiple senses were labeled by emotions related to a secondary affective sense, not their main non-affective sense. I manually reviewed and removed these incorrect labels. Additionally, I identified missing lemma-POS pairs with the help of closely related pairs already labeled by WNA. For example the pair *glorious-JJ* was missing from WNA, but is related (via the *derived-from* relation) to already labeled pair *glorify-VB*. I manually searched for these missing relationships, adding the missing terms, as well as recursively adding their synonyms (e.g., *glorious-JJ* resulted in *splendid*, *magnificent*, *brilliant*, and *superb* being added as well). In total, I removed 613 and added 814 labels of different lemma-POS pairs, resulting a final count of 4048 lemma-POS pairs.

In general, the technique of using a fixed lexicon of emotion terms to capture highly context-dependent emotional expressions is problematic at best. Although I show here that work on improving the lexicon does improve emotion recognition results, ultimately, any technique will have to move away from a rigid lexicon-based approach to something more flexible. I'll go through these topics in more detail in the following chapters.

## Step 1: Pre-Processing

For each sentence $s \in S$ in the given corpus, I construct a bag of words by tokenizing the sentence and lemmatizing each word. I generate a count vector for $\text{BoW}_s$ by mapping each lemma to the count in the sentence ($\Omega \mapsto \mathbb{Z}_{\geq 0}$). I do not remove stop words as their effects are minimized by the *tf-idf* computation in the next step.

I define the bag of words of each sentence as a mapping of its words to their occurrence frequency in each sentence; therefore, for any sentence $s$ and word $\omega$, $\text{BoW}_s(\omega) > 0$, iff $\omega$ participates in $s$. Then, I create the set $T$ of all lemmatized terms of the whole corpus in the following way:

$$T = \left\{ t \in \Omega : \exists s \in S \text{ s.t. } \text{BoW}_s(t) > 0 \right\} \tag{3.1}$$

I use $m$ to represent the total number of terms in $T : t_1, t_2, \ldots, t_m$.

## Step 2: Vector Space Modeling

Using the count vectors constructed in the first step, I compute a *tf-idf* vector for each sentence as well as a standard vector for each emotion class $e \in E$. For each sentence $s_j \in S$, I construct an $m$ dimensional vector where each entry in the vector is the *tf-idf* of term $\omega_i$ in sentence $s_j$; i.e.

$$v_{ij} = \text{TF}_{i,j} \times \text{IDF}_i \tag{3.2}$$

where $\text{TF}_{i,j} = \text{BoW}_{s_j}(\omega_i)$,

$$\text{IDF}_i = \log \frac{n}{|\{s \in S : BoW_s(\omega_i) > 0\}|}. \tag{3.3}$$

$n$ is the number of sentences, and $\Omega = \{\omega_i\}_{i=1}^{m}$.

The constructed vector space model is represented by the following $m \times n$ matrix $V$:

$$V = [V_{s_1} V_{s_2} \dots V_{s_n}] \text{ where } V_{s_j} = \begin{pmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{mj} \end{pmatrix} \tag{3.4}$$

I compute a standard vector for each emotion class $Y_e = (y_{e,\omega_1}, y_{e,\omega_2}, \dots, y_{e,\omega_m})$ where $y_{e,\omega_i}$ is 1 if the term $\omega_i$ is mapped to $e$ by the lexicon, otherwise 0.

## Step 3: Noise Cancellation or Dimension Reduction

The vectors $V_s$ and $Y_e$ from the previous step are all $m$-dimensional vectors where $m$ is the total number of terms in the corpus. There are many terms that have little or no effect on the emotion labeling of their sentences. Therefore, dimensional reduction or noise cancellation techniques may improve the performance of the emotion labeling step which comes later. Principle Component Analysis (PCA) has been known for quite some time for noise cancellation [Abdi and Williams, 2010], while Latent Dirichlet Allocation (LDA) was specifically developed for dimension reduction in natural language processing [Blei et al., 2003]. Non-Negative Matrix Factorization (NMF) was first introduced for noise cancellation by Lee and Seung [1999].

### Step 3.1: Vector Space Decomposition

I can decompose the obtained matrix $V$ in one of the following three ways:

1. Non-negative Matrix Factorization (NMF): I extract $d$ features from the $m$-dimensional vectors of sentences using NMF.
2. Principal Component Analysis (PCA): I reduce the number of dimensions of $V_s$ vectors from $m$ to $\Delta < m$.

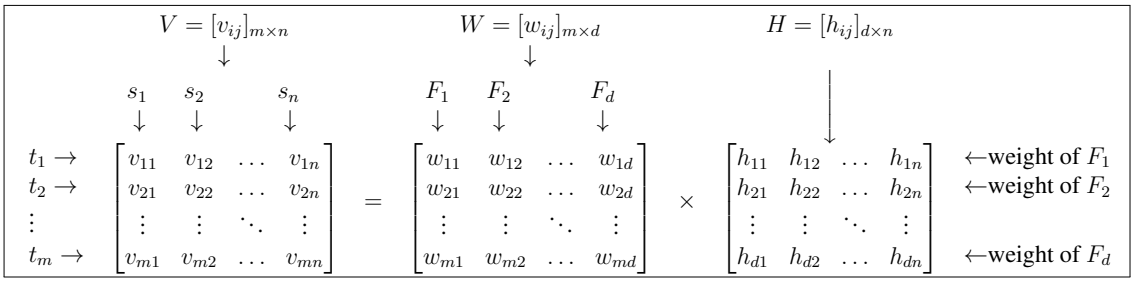



Figure 3.2: Non-negative matrix factorization (Step 3.1) to extract features of sentence vector model $V$. The results of this process is given by matrices $W$ and $H$. Columns of $W$ are corresponding to the extracted features $F_1, F_2, \ldots, F_d$ of the model and rows of $H$ are called the weights of these features.

3. Latent Dirichlet Allocation (LDA): I reduce the number of dimensions of $V_s$ vectors from $m$ to $\delta < m$.

When using PCA or LDA I can move directly to fourth step of the system; however, in the case of NMF, I must select important terms (Step 3.2), remove irrelevant features (Step 3.3), and reconstruct the vector space (Step 3.4).

When using NMF for decomposing the vector space model, $V$ is factorized into two matrices $W_{m\times d} = [w_{ij}]$ and $H_{d\times n} = [h_{ij}]$, both with all non-negative entries:

$$V = W \times H \text{ s.t. } w_{ij} \geq 0 \text{ and } h_{ij} \geq 0 \tag{3.5}$$

Note that $d$ is considered a hyper-parameter in this step and its numerical value can be fine-tuned by maximizing the output of the system on a development set.

The NMF factorization process produces a matrix $W$ whose $d$ columns each represents an $m$-dimensional feature for each of the original $n$ sentences in the corpus:

$$W = [F_1 F_2 F_3 \ldots F_d] \text{ where } F_j = \begin{pmatrix} w_{1j} \\ w_{2j} \\ \vdots \\ w_{mj} \end{pmatrix} \tag{3.6}$$

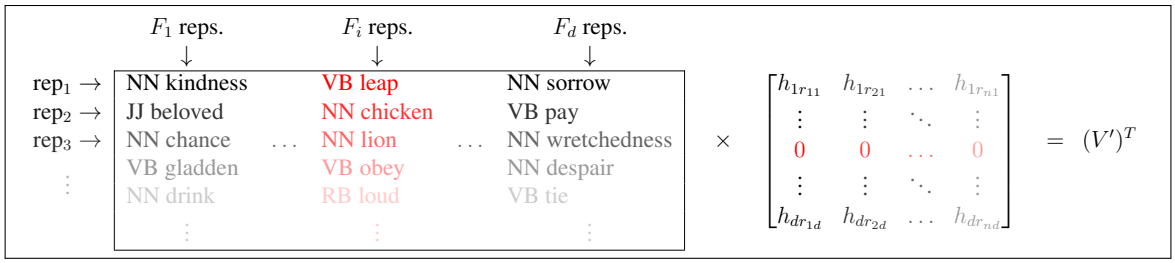
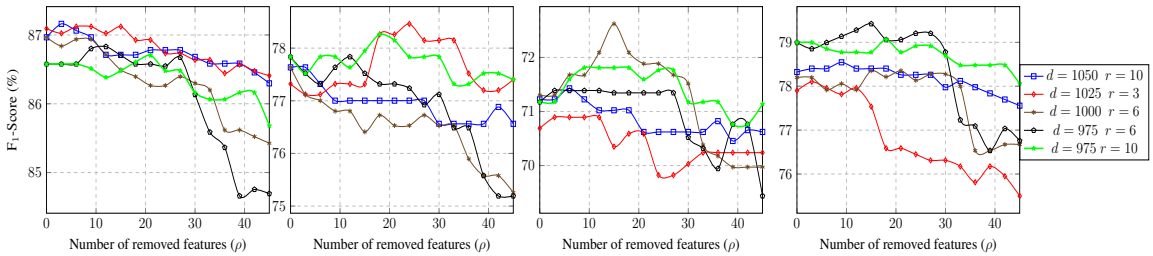



Figure 3.3: Vector space reconstruction. The least relevant features are removed by zeroing out their corresponding weights in matrix $H$. The updated $H$ matrix is denoted by $H'$. The sentence vector model is then reconstructed by multiplying $W$ by $H'$ (Steps 3.3 & 3.4). The updated sentence vector model is represented by matrix $V'$.




(a) Joy (b) Anger (c) Fear (d) Sadness

Figure 3.4: Scores of various setups of the proposed model using NMF. Each combination of hyper-parameters $d$, $r$, and $\rho$ (dimensions, representatives, and removed features) results in a specific $F_1$ score for each emotion label. The model with $(d, r) = (975, 10)$, highlighted with green color, results in the highest overall $F_1$ score when $\rho = 18$. For each individual emotion, the best $F_1$ score is found at (a) Joy: $(d, r, \rho) = (1050, 10, 3)$, (b) Anger: $(d, r, \rho) = (1025, 3, 24)$, (c) Fear: $(d, r, \rho) = (1000, 6, 15)$, (d) Sadness: $(d, r, \rho) = (975, 6, 15)$.

Each of the $d$ rows of $H$ matrix represents weights of the $d$ features in $F$. This decomposition is shown in Figure 3.2.

### Step 3.2: Term Selection

For every feature $F_j$, I identify a fraction $r$ of terms with the highest weights as its representatives, where $r$ is a hyper-parameter that can be fine-tuned during system optimization ($r$ is usually less than 1%).

### Step 3.3: Feature Removal

In this phase I remove the $\rho$ features that have little or no emotional relevance, where $\rho$ is a non-negative integer hyper-parameter that can be tuned. I will call a feature "emotionally irrelevant" if all of its representative terms (as selected in the previous step) are labeled as neutral by the lexicon. These features will always be removed first. If $\rho$ is less than the number of emotionally irrelevant features, I choose at random. On the other hand, if the number of emotionally irrelevant features is less than $\rho$, I eliminate features $F_j$ in order of their overall emotional relevance, which is computed by estimating the standard deviation of cosine similarity ratios between emotion vectors $Y_e$'s obtained in Step 2 and $F_j \circ R_j$ (element-wise product of $F_j$ and $R_j$) where $R_j$ is the binary identifier of whether a term is a representative for $F_j$ and is constructed based on the outcome of Step 3.2. Symbolically, to quantify how emotionally relevant feature $F_j$ is, I calculate the following standard-deviation:

$$\sigma_j = \text{StdDev}_{e \in E \setminus \text{neutral}} \big\{ \text{sim}_{\cos}(Y_e, F_j \circ R_j) \big\} \tag{3.7}$$

### Step 3.4: Vector Space Reconstruction

In this step, the vector space model is reconstructed ($V'$) after eliminating the irrelevant features. Let $I$ denote the set of indices whose corresponding features are identified as least relevant in previous step. Then the reconstructed vector space is:

$$V' = [v'_{ij}]_{m \times n} \text{ s.t. } v'_{ij} = \sum_{\substack{1 \le k \le d \\ k \notin I}} w_{ik} h_{kj} \tag{3.8}$$

Figure 3.3 illustrates the vector space reconstruction.

### Step 4: Labeling

Finally the emotion recognition process takes place by measuring the similarity between sentence vectors $V_s$ and standard emotion vectors $Y_e$ which are taken from the previous step with the help of NMF, PCA, or LDA. Label of each sentence $s$ is calculated by the following formula:

$$\text{predicted label of } s = \arg\max_{e \in E} \ \text{sim}(V_s, Y_e) \tag{3.9}$$

where similarity function can be measured by the cosine of angle made by the two given vectors:

$$\text{sim}_{\cos}(V_s, Y_e) = \frac{V_s \cdot Y_e}{||V_s|| \times ||Y_e||} \tag{3.10}$$

## 3.3 Performance on Fairy Tale Data

I tuned and tested my system using the manually annotated dataset of fairy tales constructed by Alm [2008], which comprises 176 children's fairy tales (80 from Brothers Grimm, 77 from Hans Andersen, and 19 from Beatrix Potter) with 15,087 unique sentences (15,302 sentences), 7,522 unique words and 320,521 total words. These fairy tales were annotated by two annotators labeling the emotion and mood of each sentence as one of joy, anger, fear, sadness, or neutral which resulted in four labels per sentence. Across the sentences, only 1,090 of them agreed on *all four non-neutral labels*. Kim et al. [2010] used only these sentence to train and test their system[1], and I followed the same procedure. There were 2,405 unique term-POS pairs. Also, the distribution of labels in the dataset is specified in the pie-chart depicted in Figure 3.5.

[1] Kim et al. [2010] reported 1,093 sentences, but I found and removed three sentences that were repeated in the data.

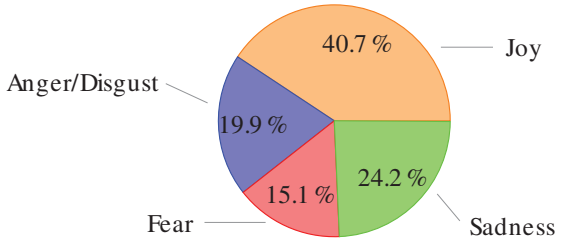



Figure 3.5: Alm's fairy tales label distribution.

| **Sentence** | **Predicted** | **Gold Label** |
|---|---|---|
| *They told him that their father was very ill, and that they were afraid nothing could save him.* | Fear | Sadness |
| *And in sight of the bridge! Said poor pigling, nearly crying.* | Sadness | Fear |
| *She smiled once more, and then people said she was dead.* | Sadness | Joy |
| *Then he aimed a great blow, and struck the wolf on the head, and killed him on the spot!* | | |
| ... and when he was dead they cut open his body, and set Tommy free. | Anger | Joy |

Table 3.1: Challenging examples of sentences incorrectly labeled by the model.

I measured the performance of my system on Alm's data. Without augmenting WNA, using the original 1,471 synsets of WNA, the $F_1$ score is 0.625. The performance metrics presented in Table 3.2 were obtained by the model using the augmented WNA. The plots depicted in Figure 3.4 show the $F_1$ scores of various setups of the proposed model using NMF technique for noise cancellation. Also, Table 3.2 summarizes the precision, recall and $F_1$ score of my system for each of the four emotion classes as well as its overall $F_1$ score when using NMF, PCA, or LDA with different setups (values of hyper-parameters). As observed in this table, the highest overall $F_1$ score is obtained when using NMF with $(d, r, \rho) = (975, 10, 18)$. In this model, 209 sentences were labeled incorrectly. Among them, some challenging examples are in Table 3.1.

The models specified in Table 3.2 have resulted in the highest $F_1$ scores among all of the models tested during the process of hyper-parameter optimization. This process has tested PCA-assisted and LDA-assisted mechanisms for dimensions $100, 150, 200, \ldots, 2400$ and NMF-assisted mechanism for number of folds $d = 100, 150, \ldots, 1050$, number of representatives $r = 1, 2, \ldots, 15$, and number of excluded features $x = 5, 10, \ldots, 55$.

| | | Joy | | | Anger | | | Fear | | | Sadness | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Setup | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $F_1$ | Acc. |
| NMF | 1050,10,3 | 0.872 | 0.872 | **0.872** | 0.878 | 0.696 | 0.776 | 0.672 | 0.758 | 0.712 | 0.753 | 0.818 | 0.784 | 0.807 | 0.806 |
| | 1025,3,24 | 0.859 | 0.876 | 0.867 | 0.884 | 0.705 | **0.785** | 0.682 | 0.715 | 0.698 | 0.733 | 0.799 | 0.764 | 0.800 | 0.799 |
| | 1000,6,15 | 0.872 | 0.858 | 0.865 | 0.861 | 0.687 | 0.764 | 0.692 | 0.764 | **0.726** | 0.742 | 0.830 | 0.784 | 0.804 | 0.803 |
| | 975,6,15 | 0.860 | 0.874 | 0.867 | 0.882 | 0.691 | 0.775 | 0.689 | 0.739 | 0.713 | 0.759 | 0.833 | **0.794** | 0.808 | 0.807 |
| | 975,10,18 | 0.858 | 0.874 | 0.866 | 0.879 | 0.705 | 0.783 | 0.703 | 0.733 | 0.718 | 0.755 | 0.830 | 0.791 | **0.809** | **0.808** |
| PCA | 1050 | 0.884 | 0.775 | **0.826** | 0.760 | 0.700 | 0.729 | 0.552 | 0.770 | 0.643 | 0.756 | 0.777 | 0.766 | **0.760** | **0.689** |
| | 1150 | 0.885 | 0.764 | 0.820 | 0.743 | 0.719 | **0.731** | 0.542 | 0.745 | 0.628 | 0.748 | 0.765 | 0.757 | 0.752 | 0.683 |
| | 950 | 0.883 | 0.766 | 0.820 | 0.722 | 0.696 | 0.709 | 0.571 | 0.782 | **0.660** | 0.759 | 0.777 | 0.768 | 0.757 | 0.686 |
| | 1100 | 0.888 | 0.768 | 0.824 | 0.744 | 0.710 | 0.726 | 0.542 | 0.745 | 0.628 | 0.765 | 0.788 | **0.776** | 0.758 | 0.684 |
| LDA | 1650 | 0.636 | 0.768 | **0.696** | 0.597 | 0.498 | 0.543 | 0.414 | 0.424 | 0.419 | 0.603 | 0.466 | 0.526 | 0.589 | 0.589 |
| | 1350 | 0.598 | 0.791 | 0.681 | 0.651 | 0.558 | **0.600** | 0.482 | 0.333 | 0.394 | 0.522 | 0.402 | 0.454 | 0.581 | 0.581 |
| | 1300 | 0.584 | 0.809 | 0.678 | 0.566 | 0.475 | 0.516 | 0.594 | 0.461 | **0.519** | 0.570 | 0.356 | 0.438 | 0.580 | 0.580 |
| | 2350 | 0.671 | 0.640 | 0.655 | 0.524 | 0.498 | 0.511 | 0.456 | 0.497 | 0.475 | 0.584 | 0.621 | **0.602** | 0.585 | 0.585 |
| | 1700 | 0.652 | 0.696 | 0.673 | 0.622 | 0.516 | 0.564 | 0.454 | 0.533 | 0.490 | 0.603 | 0.553 | 0.577 | **0.601** | **0.601** |

Table 3.2: Comparison of different models for detecting different emotions. The upper part of the table shows performance of the proposed model using NMF technique with different values of $(d, r, \rho)$; while the middle and bottom parts determine the model accuracy when PCA and LDA techniques are used respectively. The highest $F_1$ scores of each noise cancellation technique are highlighted.

## 3.4 Unsolved Challenges and Future Work

As already discussed, one challenge regarding automatic emotion recognition is the context dependency of emotional semantics. For instance, *I'm over the moon!* is an expression of extreme happiness but does not use any explicitly happy or joyful words (or, indeed, any emotion word at all). Another obstacle is polysemous words, when words have both an emotional and non-emotional senses; recognizing which sense of the word is being used is challenging and remains an open problem. Aside from these fundamental issues, there is a serious lack of high-quality annotated data, not just for narrative text but for all discourse types. Annotated corpora use a wide variety of sometimes incompatible emotion theories and are often poorly annotated, with low inter-annotator agreements and many errors.

There are many remaining challenges in the context of emotion extraction from textual data. One of the important tasks is to fine-tune the existing emotion detection pipeline by finding out which combination of machine learning techniques and dimension reduction

methods results in the best performance of emotion extraction. Different machine learning techniques can be used to perform emotion labeling. Given these considerations, there are many possible directions for future work, for example:

- Reconciling emotion lexicons and context dependency of emotion detection models using learning techniques;
- Evaluating the performance of a bag-of-words multi-layer perceptron applied to the dataset to extract emotions;
- Applying multi-label prediction to the dataset and comparing the results with this work,
- Evaluating the effect of text unit size (sentence, paragraph, story) on the accuracy of sentiment labels; i.e., would there be an advantage in grouping sentences into longer units (e.g. paragraphs) and assigning a single label to this longer unit? It seems that a sentence by itself might not always carry sufficient cues to disambiguate its emotion, but its surrounding sentences might give this context.

## 3.5 Contributions

I identified a high performing approach to emotion recognition in narrative text [Kim et al., 2010] and carefully reimplemented and characterized the technique, exploring a design space of three different noise cancellation or dimension reduction techniques (NMF, PCA, or LDA), exploring various hyper-parameter settings. my experiments indicated that NMF performed best, with an overall $F_1$ of 0.809. In the course of my investigation I clarified numerous implementational issues of the work reported by Kim et al. [2010], as well as made some improvements to WordNet Affect (WNA), one of the language resources used in the system, by adding new terms manually and using Wordnet similarity relations. This work suggests several promising future directions for improving the work,

including careful annotation of a larger corpus, and augmenting WNA or similar lexicons to provide improved coverage of emotion terms. I release my code and data to enable future work[2].

[2]Code and data can be downloaded from `https://doi.org/10.34703/gzx1-9v95/03RERQ`

Chapter 4

**Revised NRC Emotion Lexicon**

## 4.1 Introduction

Emotion detection is an NLP task that has long been of interest to the field [Hancock et al., 2007, Danisman and Alpkocak, 2008, Agrawal and An, 2012], and is usually conceived as a single- or multi-label classification in which zero (or more) emotion labels are assigned to variously defined semantic or syntactic subdivisions of the text. The importance of this task has only grown as the amount of available affective text has increased: social media, in particular, has made it especially convenient for people around the world to express their feelings and emotions regarding events large and small. Moreover, due to recent advancements of e-commerce and online shopping, the number of product and service reviews, which mostly carry the feelings of buyers toward a specific product or service. Also, during the past couple of years, there has been a great shift from traditional in-person to e-learning modalities in various levels of educational institutions. Many of the face-to-face interactions among learners and instructors have been then replaced by digital written interactions in the form of public forum participation, chatting, and exchange of messages and emails. Such great volume of data retrieved from social media, online shopping review database, and educational platforms have made it possible to build and train various classification models which may rely on both machine-learning models and emotion lexicons.

feelings, ideas, and thoughts which may carry their feelings about everyday events, which happen in their individual or social lives. There are generally two ways to express emotions in textual data [Al-Saqqa et al., 2018]. First, emotions can be expressed using *emotive* vocabulary: words directly referring to emotional states (*surprise*, *sadness*, *joy*). Second, emotions can be expressed using *affective* vocabulary: words whose emotional

content depends on the context, without direct reference to emotional states, for example, interjections (*ow!*, *ouch!*, *ha-ha!*).

An *emotion lexicon* is a specific type of linguistic resource that maps the emotive or affective vocabulary of a language to a fixed set of emotion labels (e.g. Plutchik's eight-emotion model), where each entry in the lexicon associates a word with zero or more emotion labels. Because this information is difficult to find elsewhere, emotion lexicons are often used as one of the key components of affective text mining systems [Yadollahi et al., 2017]. However, as is usual with linguistic resources, creating an emotion lexicon is a time-consuming, costly, and sometimes impractical part of the task. The difficulty is only accentuated when one considers the many affective uses of words, in which the emotional content is context dependent. Such context dependency underlines the utility of General-Purpose (context-independent) Emotion Lexicons (GPELs), which captures the mostly fixed emotive content of words, and which can serve as a foundation for more context-dependent systems.

In this chapter, I analyze and improve one of the most commonly used GPELs, namely, the NRC lexicon [Mohammad et al., 2013, Mohammad and Turney, 2013, 2010]. The NRC used Macquarie's Thesaurus [Bernard, 1986] as the source for terms, retaining only words that are repeated more than 120,000 times in Google n-gram corpus [Michel et al., 2011]. The NRC maps each word to zero or more labels drawn from Plutchik's 8-emotion psychological model [Plutchik, 1980], and provides labels for 14,182 individual words. NRC is created through Amazon's Mecahanical Turk[1] and Roget's Thesaurus[2]. The Google n-gram corpus, which is available through the Linguistic Data Consortium[3], was used to annotate it [Mohammad et al., 2013].

[1]http://www.mturk.com/mturk/welcome

[2]http://www.gutenberg.org/ebooks/10681

[3]https://www.ldc.upenn.edu

While the NRC has been used extensively across the emotion mining literature [Tabak and Evrim, 2016, Abdaoui et al., 2017, Rose et al., 2018, Lee et al., 2019, Ljubešić et al., 2020, Zad et al., 2021d], close inspection reveals a large number of incorrect, non-sensical, pejorative, or otherwise troubling entries. While I provide more examples later in this chapter, to give a flavor of the problem, the NRC provides emotion labels for many generic nouns (*tree*→ANGER), common verbs (*dance*→TRUST), colors (*white*→ANTICIPATION), places (*mosque*→ANGER), relations (*aunt*→TRUST), and adverbs (*scarcely*→SADNESS). Furthermore, the NRC suffers from significant ambiguity because it does not include part of speech categories for the terms: for example, while *console* implies SADNESS in its most common verb sense (as the NRC indicates), in its most common noun sense means a small side table, which probably should have no emotive content. In my analysis, many of these problematic entries seem to stem from a conflation of *emotive* (context-independent) and *affective* (context-dependent) emotion language use: it is as if, during the annotation of Shakespeare's *Macbeth*, the annotators of the NRC marked *hell*→ANGER and *woman*→ANGER because of the bard's highly contextualized statement "Hell hath no fury like a woman scorned": while it is true that this statement is often cited to support an assertion that women are angry people in general, and such a lexicon entry would help in correct marking of the affective implication of this specific statement in this particular context, it does not generalize to all, or even most, uses of the word *woman*. Therein lies the rub.

I discuss in detail the deficiencies of the NRC, giving a variety of problematic examples, and speculating as to how these entries were included (§4.2). Next I describe a semi-automatic procedure designed to filter out many of these deficiencies (§4.3), after which I evaluate the effectiveness of the filtering procedure by integrating the corrected version of the NRC into an emotion detection system (§4.4). I conclude with a list of my contributions (§6.7).

## 4.2 Problems with the NRC

In my close inspection of the entries in the NRC, I noted three main problems. First, the NRC does not indicate the part of speech of terms labeled with emotion. This obviously causes a great deal of ambiguity as to whether a particular emotion label should apply to a particular use of a word form. Second, the NRC contains numerous incorrect, inaccurate, nonsensical, or pejorative associations, most of which can be ascribed to an apparent conflation of the distinction between emotive and affective emotional language, i.e., ignoring the importance of context for emotional semantics. Third, and finally, there are emotion markings in the lexicon for which I can find no support in Keyword-in-Context (KWIC) databases for any sense; I count these as simple errors.

### 4.2.1 Missing Parts of Speech

As Mohammad and Turney [2010] noted, the NRC includes some of the most frequent English nouns, verbs, adjectives, and adverbs. Problematically, however, the NRC does not indicate the part of speech for any entry. For example, the wordform *bombard* is labeled as ANGER|FEAR; however, in WordNet the gloss for the first sense of *bombard* as a noun is "a large shawm[4] ; the bass member of the shawm family". On the other hand, the gloss of the first sense of the verb form of *bombard* is "cast, hurl, or throw repeatedly with some missile", which is more compatible with the emotion ANGER|FEAR. Another example is the word *console*. The NRC marks *console*→SADNESS, but the primary sense of the noun form refers to "a small table fixed to a wall or designed to stand against a wall." Clearly there is no context-independent emotional inflection to this sense. The

[4]a *shawm* is a type of musical intstrument

SADNESS label is more appropriate for the first verb sense "to give moral or emotional strength to", usually to a sad person.

Despite Araque et al. [2019] claims that adding POS tags to lexicons may decrease the performance of emotion detection mechanisms, I observe that lack of POS tagging has caused considerable ambiguity which negatively affects my emotion detection system performance. Mixing terms and labeling them without assigning pos-tags makes the lexicon less accurate as there are many words that will get different senses and emotion labels when they are assigned to different pos-tags.

Table 4.1 lists a small selection of NRC wordform labels that are problematic because of part-of-speech-related ambiguity. Although I did not count the number of NRC entries suffering this particular part-of-speech ambiguity problem, my best guess is that it affects roughly several thousand entries, about a third of the non-neutral portion of the lexicon. In this work I assigned part of speech tags to all the NRC words that exist in WordNet and label them based on the first sense definition of the terms in WordNet.

### 4.2.2 Context Dependency

In general-purpose emotion lexicons (GPELs), words are generally marked with an emotion (one or more labels) if there is a dominant sense of the word, and it has emotion semantics. In domain-specific emotion lexicons (DSELs), by contrast, assignment of an emotion label is based on the common sense of each term in a specific domain [Bandhakavi et al., 2017]. For example, the noun "shot" in a DSEL tailored for *sports*, referring taking a shot at a goal, might be plausibly marked as (*shot*→ANTICIPATION|JOY), while in a medical DSEL, referring to a injection, might be marked as (*shot*→ANTICIPATION|FEAR). Similarly, the adjective "crazy" in sports might be marked according to the sense in the statement "that goal was crazy!" (*crazy*→JOY|SURPRISE) while in the behavioral do-

| Word | POS | Original NRC Labels | First Sense in WordNet | Corrected Label |
|---|---|---|---|---|
| accidental | NN | FEAR\|SURPRISE | a musical notation that makes a note sharp or flat or natural | NEUTRAL |
| account | NN | TRUST | a record or narrative description of past events | NEUTRAL |
| ail | NN | SADNESS | aromatic bulb used as seasoning | NEUTRAL |
| alien | NN | DISGUST\|FEAR | a person who comes from a foreign country | NEUTRAL |
| allure | VB | ANTICIPATION\|JOY\|SURPRISE | dispose or incline or entice to | DISGUST\|ANTICIPATION |
| award | NN | ANTICIPATION\|JOY\|SURPRISE\|TRUST | a grant made by a law court | SADNESS |
| awful | RB | ANGER\|DISGUST\|FEAR\|SADNESS | used as a verbal intensifier | NEUTRAL |
| baby | NN | JOY | a very young child (birth to 1 year) who has not yet begun to walk or talk | NEUTRAL |
| bad | RB | ANGER\|DISGUST\|FEAR\|SADNESS | with great intensity ( ‘ bad ’ be a nonstandard variant for ‘ badly ’ ) | NEUTRAL |
| badger | NN | ANGER | a native or resident of Wisconsin | NEUTRAL |
| bark | NN | ANGER | tough protective covering of the woody stems and roots of trees and other woody plants | NEUTRAL |
| batter | NN | ANGER\|FEAR | (baseball) a ballplayer who is batting | NEUTRAL |
| bayonet | NN | ANGER\|FEAR | a knife that can be fixed to the end of a rifle and used as a weapon | NEUTRAL |
| beam | NN | JOY | a signal transmitted along a narrow path | NEUTRAL |
| belt | NN | ANGER\|FEAR | endless loop of flexible material between two rotating shafts or pulleys | NEUTRAL |
| bias | JJ | ANGER | slanting diagonally across the grain of a fabric | NEUTRAL |
| blister | NN | DISGUST | a flaw on a surface resulting when an applied substance does not adhere | NEUTRAL |
| blitz | NN | SURPRISE | (American football) defensive players try to break through the offensive line | NEUTRAL |
| bloody | RB | ANGER\|DISGUST\|FEAR\|SADNESS | extremely | NEUTRAL |
| blossom | NN | JOY | reproductive organ of angiosperm plants especially one having showy or colorful parts | NEUTRAL |
| board | VB | ANTICIPATION | get on board of (trains, buses, ships, aircraft, etc.) | NEUTRAL |
| boil | VB | DISGUST | come to the boiling point and change from a liquid to vapor | NEUTRAL |
| bombard | NN | ANGER\|FEAR | a large shawm ; the bass member of the shawm family | NEUTRAL |
| boomerang | NN | ANTICIPATION\|TRUST | a curved piece of wood; when properly thrown will return to thrower | NEUTRAL |
| buffet | NN | ANGER | a piece of furniture that stands at the side of a dining room; has shelves and drawers | NEUTRAL |
| bully | JJ | ANGER\|FEAR | very good | SURPRISE\|JOY |
| cage | NN | SADNESS | an enclosure made or wire or metal bars in which birds or animals can be kept | NEUTRAL |
| case | NN | FEAR\|SADNESS | an occurrence of something | NEUTRAL |
| chaff | NN | ANGER\|FEAR | material consisting of seed coverings and small pieces of stem or leaves | NEUTRAL |
| collateral | JJ | TRUST | descended from a common ancestor but through different lines | NEUTRAL |
| connective | NN | TRUST | an uninflected function word that serves to conjoin words | NEUTRAL |
| console | NN | SADNESS | a small table fixed to a wall or designed to stand against a wall | NEUTRAL |
| desert | NN | ANGER\|DISGUST\|FEAR\|SADNESS | arid land with little or no vegetation | NEUTRAL |
| kind | NN | JOY\|TRUST | a category of things distinguished by some common characteristic or quality | NEUTRAL |
| present | NN | ANTICIPATION\|JOY\|SURPRISE\|TRUST | the period of time that is happening now; | NEUTRAL |
| sentence | NN | ANTICIPATION\|DISGUST\|ANGER\|FEAR\|SADNESS | a string of words satisfying the grammatical rules of a language | NEUTRAL |
| rail | NN | ANTICIPATION\|ANGER | a barrier consisting of a horizontal bar and supports | NEUTRAL |

Table 4.1: Examples of NRC terms with inappropriate emotion labels and correction. The last column shows the proposed correction.

main, it might be (*crazy*→DISGUST|FEAR). Table 4.2 gives a small selection of NRC entries where each label is appropriate only in a limited context, not corresponding to the literal meaning of the word in its dominant sense. The extreme version of this problem can be seen with words like *abundance* which have a multitude of labels that conflict (DISGUST|JOY|TRUST|ANTICIPATION). Overall this is a problem with regards to NRC because it is explicitly presented as a GPEL. In my evaluation of the NRC, while again I did not count exactly how many entries suffered from this issue, I estimate at least 600 or so entries, or 10% of the NRC, fall into this category.

### 4.2.3 Simple Errors

The NRC has a large number of terms, and as with any resource of this size there are bound to be minor faults or errors. Since human annotators provided the data needed to

| Term | NRC Labels | Term | NRC Labels |
|---|---|---|---|
| abundance | DISGUST\|JOY | monk | TRUST |
| | \|TRUST\|ANTICIPATION | oblige | TRUST |
| baby | JOY | recreation | JOY\|ANTICIPATION |
| count | TRUST | remedy | JOY |
| create | JOY | remove | ANGER\|FEAR |
| explain | TRUST | | \|SADNESS |
| fact | TRUST | saint | ANTICIPATION\|JOY |
| fall | SADNESS | | \|TRUST\|SURPRISE |
| fee | ANGER | save | JOY |
| fire | FEAR | score | ANTICIPATION\|JOY |
| gain | JOY\|ANTICIPATION | | \|SURPRISE |
| grow | ANTICIPATION\|JOY\|TRUST | star | ANTICIPATION\|JOY |
| larger | DISGUST\|SURPRISE\|TRUST | | \|TRUST |
| leader | TRUST | understand | TRUST |
| mate | TRUST | unnatural | DISGUST\|FEAR |

Table 4.2: Examples of context dependency

create the resource, I can assume that certain terms were given labels that are not appropriate and that some small number of these errors would have escaped notice of any manual error correcting procedures. I define these sorts of errors as those where the provided emotional labels do not make sense in any context supported by Keyword-in-Context (KWIC) indicies [iWeb, 2021, Davies and Kim, 2019]. Table 4.3 lists a small selection of examples of seemingly simple errors in labels, for example *architecture*→TRUST. Some markings, furthermore, might be reflective of relatively obvious biases, which in light of recent work demonstrating the built-in biases of various AI and NLP resources [Bolukbasi et al., 2016, Bender and Friedman, 2018, Mehrabi et al., 2019, Blodgett et al., 2020], it would be good to try to correct for. Examples of the latter case include the entries *fat*→DISGUST |SADNESS, *lesbian*→DISGUST|SADNESS, or *mosque*→ANGER. I estimate that the number of entries affected by simple errors or biases is at least a few hundred, or roughly 5% of the NRC.

### 4.2.4 Problems with the NRC Annotation Process

Some aspects of the NRC annotation process go part of the way toward explaining some of the above problems. As discussed by Mohammad and Turney [2013], the annotation process relied upon approximately 2,000 native and fluent speakers of English who answered a series of questions regarding the emotion terms. The directions were made ambiguous on purpose to minimize biasing the subject's judgements. The concern with this method is that the annotators could have been shown a term that is not familiar to them. This was circumvented by asking the individual to associate the term with a certain word similar in meaning amongst three non-viable options.

After selecting the most similar word, the annotator could continue annotating even when they do not really know the meaning of a word. This could have happened by the annotator quickly looking up the definition online. The annotators were told not to look up the words[5], but there is no guarantee that they did so, and much work has shown that crowdworkers are often unreliable [Ipeirotis et al., 2010, Vuurens et al., 2011].

Another concern with the annotation process was question wording. Questions 4–11 in particular raise specific concerns. These asked, for all combinations of a term $X$ and each of the eight emotions $Y$, "How much is $X$ associated with the emotion $Y$?" Posing this in only the positive formulation potentially biased annotators to find confirmatory evidence. A more balanced procedure would have been to ask annotators to imagine not only how much of emotion $Y$ was associated $X$, but also how much $Y$ *wasn't* associated with $X$, prompting them to consider disconfirmatory evidence. Because of this confirmation bias in the collection procedure I posit that many of the terms in the NRC were associated with particular emotions even when those terms do not bring those emotions to mind when mentioned in isolation in normal usage.

[5] Annotators were instructed "please skip HIT if you do not know the meaning of the word"

We created Mechanical Turk HITs for each of the terms specified in Table 1. Each HIT has a set of questions, all of which are to be answered by the same person. We requested five different assignments for each HIT (each HIT is to be annotated by five different Turkers). Different HITS may be attempted by different Turkers, and a Turker may attempt as many HITs as they wish. Below is an example HIT for the target word "startle".

**Title:** Emotions evoked by words
**Reward per HIT:** $0.04
**Directions:** Return HIT if you are not familiar with the prompt word.

Prompt word: **startle**

1. Which word is closest in meaning (most related) to *startle*?
   - automobile
   - shake
   - honesty
   - entertain

2. How positive (good, praising) is the word startle?
   - startle is not positive
   - startle is weakly positive
   - startle is moderately positive
   - startle is strongly positive

3. How negative (bad, criticizing) is the word startle?
   - startle is not negative
   - startle is weakly negative
   - startle is moderately negative
   - startle is strongly negative

4. How much does the word *startle* evoke or produce the emotion joy (for example, *happy* and *fun* may strongly evoke joy)?

Figure 4.1: Questions in Mechanical Turk Hits for each terms. Questions 5–11 repeat Question 4 but for the other seven emotions. [Mohammad and Turney, 2010]

Another way of addressing this bias would have been to show words in specific contexts; this avoids the need for an annotator to think up their own evidence to support their label, which may have been limited by the annotators's time, attention, creativity, or knowledge of English usage. Such an approach would no doubt have been costlier, but it perhaps would have produced higher quality labels.

When it came to validating the NRC, the authors compared their crowdsourced labels with labels from the WNA lexicon to see how close the judgements were. In the one earlier paper [Mohammad and Turney, 2013], when the NRC had 10,000 entries, the authors reported that only 6.5% of the entries could be matched with those in WNA.

| Term | Labels | Term | Labels |
|---|---|---|---|
| abacus | TRUST | cabinet | TRUST |
| alb | TRUST | calculation | ANTICIPATION |
| ambulance | FEAR\|TRUST | coyote | FEAR |
| ammonia | DISGUST | critter | DISGUST |
| anaconda | DISGUST\|FEAR | crypt | FEAR\|SADNESS |
| aphid | DISGUST | fat | DISGUST\|SADNESS |
| archaeology | ANTICIPATION | fee | ANGER |
| architecture | TRUST | iron | TRUST |
| assembly | TRUST | lamb | JOY\|TRUST |
| association | TRUST | mill | ANTICIPATION |
| asymmetry | DISGUST | mountain | ANTICIPATION |
| atherosclerosis | FEAR\|SADNESS | mosque | ANGER |
| baboon | DISGUST | machine | TRUST |
| backbone | ANGER\|TRUST | organ | ANTICIPATION\|JOY |
| balm | ANTICIPATION\|JOY | pine | SADNESS |
| basketball | ANTICIPATION\|JOY | rack | SADNESS |
| bee | ANGER\|FEAR | ravine | FEAR |
| belt | ANGER\|FEAR | ribbon | ANTICIPATION\|JOY |
| bier | FEAR\|SADNESS | | \|ANGER |
| biopsy | FEAR | rod | TRUST \|FEAR |
| birthplace | ANGER | spine | ANGER |
| blackness | FEAR\|SADNESS | stone | ANGER |
| bran | DISGUST | title | TRUST |
| infant | ANTICIPATION | tree | ANTICIPATION\|JOY |
| | \|FEAR\|JOY | | \|DISGUST\|TRUST |
| | \|SURPRISE | | \|SURPRISE\|ANGER |

Table 4.3: Examples of simple errors.

Later, when the NRC was expanded to 14,182 entries, the authors did not report the percentage overlap. I measured this myself, and found the overlap between the full NRC and WNA is 2,328 (16%). This is a concern because this means most of the data could not be independently validated to see how accurate the annotations were, and so a majority were not subject to any rigorous or systematic quality control check.

## 4.3 Semi-Automatic Correction of the NRC

The NRC includes 14,182 entries made up of a unigram (single token wordforms) associated with a selection of Plutchik's emotions eight (SADNESS, JOY, FEAR, ANGER, SURPRISE, TRUST, DISGUST, and ANTICIPATION), NEUTRAL, and two sentiments; as noted,

| Term | Label | Term | Label | Term | Label |
|---|---|---|---|---|---|
| arm | NEUTRAL | audience | NEUTRAL | baby | NEUTRAL |
| belt | NEUTRAL | diversity | NEUTRAL | economy | NEUTRAL |
| fat | NEUTRAL | fee | NEUTRAL | gate | NEUTRAL |
| office | NEUTRAL | buy | NEUTRAL | endpoint | NEUTRAL |
| letter | NEUTRAL | measure | NEUTRAL | money | NEUTRAL |
| rail | NEUTRAL | road | NEUTRAL | score | NEUTRAL |
| ship | NEUTRAL | star | NEUTRAL | store | NEUTRAL |
| sun | NEUTRAL | tree | NEUTRAL | word | NEUTRAL |
| clothes | NEUTRAL | filter | NEUTRAL | yeast | NEUTRAL |

Table 4.4: Examples of neutral words

no words had part of speech tags. After removing 9,719 wordforms marked neutral, examples of which are shown in Table 4.4, 4,463 wordforms remained. In the remainder of this chapter I refer to this set as `NRC.orig`. I developed a procedure to semi-automatically correct the problems discussed in prior section. First, I assigned part-of-speech tags to entries. Second, I developed an automatic emotional word test leveraging both the original version of WNA and the larger WordNet resource. Finally, I manually checked all entries for correctness. Here I provide more examples of troubling entries in the NRC that were wrongly labeled neutral words.

- nouns (e.g., tree, word, store, sun, star, audience, economy, money, fee, food, gate, measure, score, baby, rail, letter, belt, ship, fat, etc.)
- verbs (e.g., watch; FEAR, pay; JOY, dance; TRUST, cover; TRUST, teach; SURPRISE)
- colors (e.g., white; ANTICIPATION, tawny; DISGUST, green; JOY|TRUST),
- places (e.g., mosque; ANGER, desert; DISGUST|ANGER|FEAR|SADNESS, farm; ANTICIPATION, hospital; TRUST|FEAR|SADNESS, school; TRUST, saloon; ANGER),
- relations (e.g., aunt; TRUST, daughter; JOY, mamma; TRUST),
- agents (author; TRUST, lawyer; DISGUST|ANGER|FEAR|TRUST, doctor; TRUST, policeman; TRUST|FEAR, butcher; DISGUST|ANGER|FEAR),
- adverbs (e.g., finally; ANTICIPATION|JOY|DISGUST|SURPRISE|TRUST, usually; TRUST, scarcely; SADNESS),

- animals (horse; TRUST, lamb; JOY|TRUST, toad; DISGUST),
- body-part (elbow; ANGER, flesh; DISGUST, jaws; FEAR, nose; DISGUST, shoulder; TRUST, stomach; DISGUST),

Also, for sensations, here are some examples that are associated with emotional labels that are inaccurate: (sweetheart; SADNESS, ambition; JOY|TRUST, hungry; ANTICIPATION, romance; FEAR, etc.)

### 4.3.1 Assigning Part of Speech to NRC words

I began by constructing an expanded list of wordforms in NRC, each associated with a valid part of speech (POS). To determine whether a POS applied to a wordform, I looking up each wordform in WordNet under each of the main open class POS tags—Verb (VB), Adjective (JJ), Noun (NN), and Adverb (RB)—so each wordform could potentially have been associated with up to four POS tags. Every wordform was present in WordNet under at least one POS. If a WordNet sense was found for a POS, I consider that a valid tags for the wordform. After this step, my list contained has 6,166 entries of wordform-POS pairs (4,463 unique wordforms). I call this set `NRC.v1`.

### 4.3.2 Emotional Word Test

In the second step, I sought to automatically determine, on the one hand, which wordform-POS pairs likely had an emotional sense (whether emotive or affective), and on the other, pairs for which I had no direct evidence of emotional semantics. To do this, I performed the following comparisons with WNA and WordNet—if any one returned true, the pair was presumed emotional; otherwise, it was marked "unknown".

1. Is the wordform-POS pair labeled as non-neutral in WNA?

2. Is the first sense of the wordform-POS pair have a synonym labeled as non-neutral in WNA?
3. Does the WordNet gloss of the first sense of the wordform-POS pair contain words that are marked as emotional in WNA?
   (a) Find the first sense in WordNet for the wordform-POS pair.
   (b) Tokenize the gloss of the first sense.
   (c) Lemmatize the gloss.
   (d) Check if the lemmas are labeled as non-neutral in WNA.

Tokenization and lemmatization were performed with `nltk` [Loper and Bird, 2002]. The above procedure identified 2,328 out of 6,166 pairs as "presumed emotional", leaving 3,838 pairs as "unknown." In the rest of this chapter, I will refer to the lexicon of 2,328 pairs "presumed affective" pairs as `NRC.v2`.

algorithm 1 provides a psuedocode procedure combining steps one and two.

With NRC entries now organized as to whether or not they are presumed emotional (according to WNA or WordNet), I proceeded to manually check all entries. I used WNA only to remove the emotion label of some NRC wordforms. Since the number of synsets in WNA is 2,328 and the number of wordforms in NRC.v1 is 6,166 there must exist many wordforms that are not associated to WNA synsets and therefore will fail the Emotional Word Test. I did not rely soly on WNA when correcting bias in NRC, as I manually annotated every wordform in `NRC.v1` regardless of its Emotional Word Test result. I performed the below checks on all 6,166 entries in `NRC.v1`. I used the Cohen's Kappa metric to assess inter-annotator agreement [Landis and Koch, 1977], which I measured as 0.928, which represents near-perfect agreement. Notably, this emotion annotation task has much higher agreement than the sentence-level annotation emotion tasks discussed in Alm's Fairy Tales dataset (Section 2; Emotion Datasets). I suspect that this is the case for at least three reasons. First, focusing on words is an easier because sentences often

**Algorithm 1:** pos expansion and emotional sense test

```
input : a wordform w from nrc
result: outputs all valid wordform+pos (w, p) pairs, marked with whether or not they are presumed emotional.
for p ∈ {VB, JJ, NN, RB} do
    if (w, p) is not in Wordnet then
        continue;
    end
    if (w, p) is labeled in WNA then
        print(w, p);
        continue;
    end
    allSynsets ← Synsets(w, p);
    firstSynset ← allSynsets[0];
    affectiveSyn ← false;
    for (s, p) ∈ firstSynset do
        if (s, p) is labeled in WNA then
            affectiveSyn ← true;
        end
    end
    if affectiveSyn is true then
        print(w, p);
        continue;
    end
    theGloss ← firstSynset.gloss;
    affectiveLem ← false;
    for (w′, p′) in tokenize(theGloss) do
        (l, p′) ←lemmatize(w′, p′);
        if (l, p′) is labeled in WNA then
            affectiveLem ← true;
        end
    end
    if affectiveLem is true then
        print(w, p);
    end
end
```

have complex emotion valence: there might multiple emotions in a sentence. Second, the NRC words that are retained at this stage are clearly emotional, they are selected to be such, and so are less emotionally ambiguous than neutral words: there are no borderline cases. Finally, I defined a clear set of procedures for identifying the emotion, which were developed during several rounds of pilot annotation, following best practice in linguistic annotation.

- **Presumed Emotional:** For each wordform-POS pair, I examined the first sense in WordNet, any labels in WNA, and the labels in `NRC.orig` to determine if they

| | NRC.v1 | | | NRC.v2 | | | NRC.v3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Emotion label | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ |
| JOY | 0.738 | 0.570 | 0.643 | 0.805 | 0.577 | 0.672 | 0.855 | 0.572 | 0.686 |
| ANGER | 0.359 | 0.253 | 0.297 | 0.347 | 0.226 | 0.274 | 0.432 | 0.240 | 0.308 |
| SURPRISE | 0.151 | 0.263 | 0.192 | 0.144 | 0.254 | 0.184 | 0.178 | 0.254 | 0.209 |
| DISGUST | 0.095 | 0.324 | 0.147 | 0.124 | 0.353 | 0.183 | 0.137 | 0.500 | 0.215 |
| FEAR | 0.407 | 0.212 | 0.279 | 0.589 | 0.200 | 0.299 | 0.535 | 0.327 | 0.406 |
| SADNESS | 0.632 | 0.417 | 0.502 | 0.661 | 0.473 | 0.552 | 0.717 | 0.451 | 0.553 |
| macro-Avg. | 0.397 | 0.340 | 0.343 | 0.445 | 0.347 | 0.361 | 0.476 | 0.391 | **0.396** |
| micro-Avg. | 0.466 | 0.408 | 0.435 | 0.510 | 0.418 | 0.460 | 0.545 | 0.435 | **0.484** |

Table 4.5: Result of using different, corrected versions of the NRC to the Zad and Finlayson [2020] emotion detection system on Alm's fairy tales.

were compatible, focusing on identified emotional words and synonyms. If there were disagreements between the WNA and `NRC.orig` I examined the Keyword-in-Context index for that POS. In cases where it was ambiguous whether `NRC.orig`, WNA, or WordNet was the correct analysis, I defaulted to `NRC.orig`. Out of 2,328 presumed emotional pairs, 1,957 were ultimately kept as having at least one emotion label.

- **Unknown:** Pairs in this group were distinguished from the Presumed Emotional group by the lack of obvious emotional words in the WordNet glosses of the pair or its synonyms. While I examined the WordNet entries for these pairs carefully, I spent more time examining the Keyword-in-Context index to look for emotional senses. Out of 3,838 unknown pairs, ultimately 1,729 were marked as having at least one emotion label.

Figure 4.2 shows the outline of the process to construct final, corrected version of the NRC, which I refer to as `NRC.v3` in the rest of this chapter.

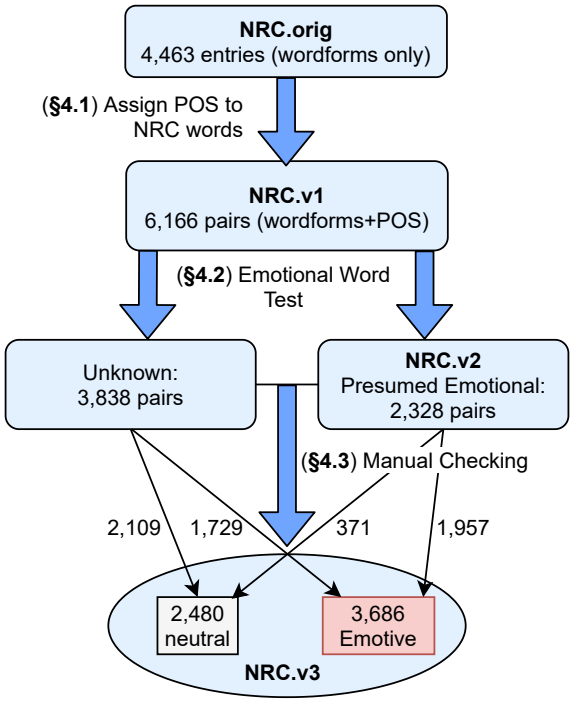



Figure 4.2: The semi-automatic procedure for correcting the NRC.

| | w SURPRISE | | | w/o SURPRISE | | | Avg. |
|---|---|---|---|---|---|---|---|
| | (1) w/ DISGUST | (2) w/o DISGUST | (3) DISGUST+ANGER | (4) w/ DISGUST | (5) w/o DISGUST | (6) DISGUST+ANGER | |
| NRC.v1 | 0.343 | 0.421 | 0.402 | 0.421 | 0.533 | 0.513 | 0.439 |
| NRC.v2 | 0.361 | 0.439 | 0.429 | 0.451 | 0.573 | 0.551 | 0.467 |
| NRC.v3 | 0.396 | 0.462 | 0.463 | 0.489 | 0.594 | 0.583 | 0.498 |
| NRC.v1 | 0.435 | 0.481 | 0.461 | 0.545 | 0.603 | 0.577 | **0.517** |
| NRC.v2 | 0.460 | 0.505 | 0.491 | 0.585 | 0.644 | 0.622 | **0.551** |
| NRC.v3 | 0.484 | 0.520 | 0.517 | 0.607 | 0.655 | 0.637 | **0.570** |

Table 4.6: Comparing the macro-average (top three rows) and micro-average (bottom three rows) $F_1$-scores of using the three corrected versions of NRC with Zad and Finlayson's emotion detection system on Alm's fairy tales using different emotion label sets.

## 4.4 Evaluation of the Corrected Resource

In order to compare and evaluate the outcome of the correction procedure, I ran the emotion detection model developed by Zad and Finlayson [2020] using `NRC.v1`, `NRC.v2`, and `NRC.v3` as the emotion lexicon. I chose this model because the code was helpfully provided in full, and the model uses a single emotion lexicon with wordform-POS pairs to drive its emotion detection. In this section, I discuss the details of this comparison.

The emotion detection system of Zad and Finlayson originally used WNA as the emotion lexicon (leveraging wordform+POS pairs), and tested on Alm's fairy tale dataset [Alm, 2008]. While the system is convenient as an experimental testbed because the full code is available, Alm's dataset uses only six emotions (ANGER, FEAR, SADNESS, SURPRISE, DISGUST, and JOY), as opposed to Plutchick's eight used by the NRC. This means I needed to trim my NRC versions down to six labels for compatibility (I dropped ANTICIPATION and TRUST). This makes the evaluation of the NRC using this experimental setup at best an approximation for the quality of my procedure. One would imagine that, if I had an experimental tested that used all eight of Plutchik's emotions, performance would be correspondingly higher.

As described below, I also experimented with reducing the number of labels, following the experimental procedure outlined in Zad and Finlayson [2020]. Further, following the same procedure, I conducted my emotion detection comparisons on the subset of Alm's dataset [Alm, 2008, 2010] which contains 15,302 sentences from 176 children's fairy tales from classic collections by Beatrix Potter, the Brother's Grimm's, and Hans C. Andersen. I chose this corpus because of the ready availability of an emotion detection system as illustrated in Figure 4.3 [Zad and Finlayson, 2020] that uses this corpus for evaluation (§3).

Alm merged anger and disgust class for data sparsity and related semantics between them in the data set and present them in one class of emotion. See Fig. 3.5 for further details.

### 4.4.1 Comparing `NRC.v1`, `NRC.v2`, and `NRC.v3`

Table 4.5 shows the precision, recall and $F_1$ measurements of the emotion detection system when substituting the three different versions of the NRC in experimental setup for

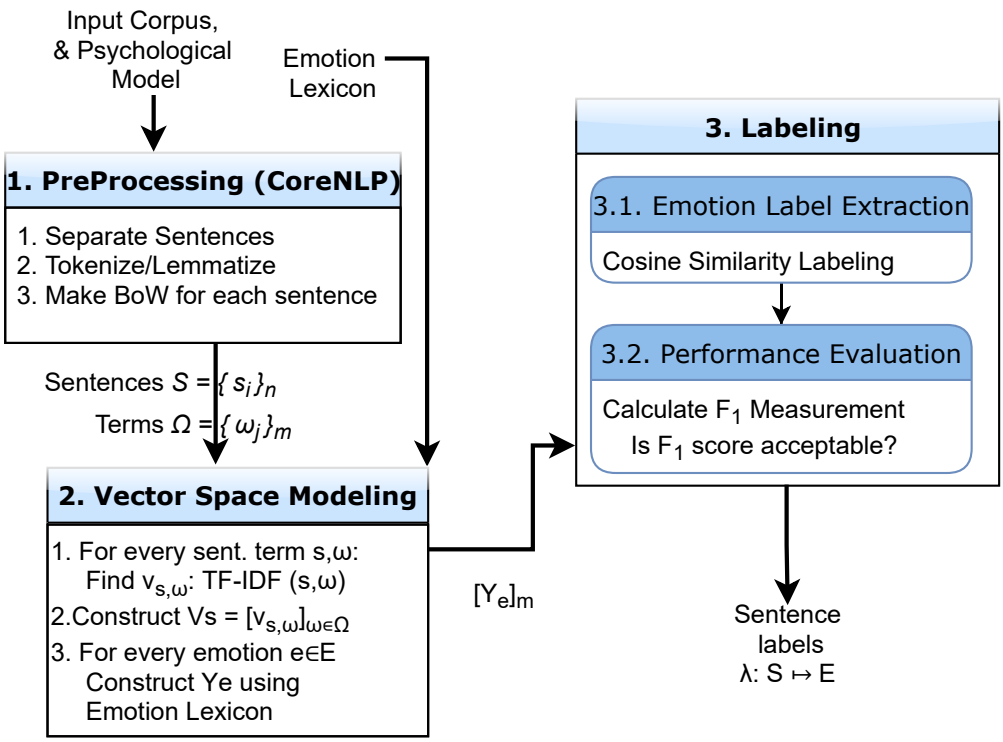



Figure 4.3: Emotion Detection System

WNA, using just the six emotions present in Alm's data (dropping all the labels of ANTICIPATION and TRUST). The first three columns result gives a baseline for performance of what is effectively the original NRC in the Zad and Finlayson [2020] experimental setup.

The next two groups show `NRC.v2` and `NRC.v3`, respectively. As can be seen, overall micro-average performance rises from 0.435 for `NRC.v1` to 0.460 for `NRC.v2` and 0.484 for `NRC.v3`. This provides solid evidence that my correction procedure improved the quality of the resource.

While one might expect that the recall in Table 4.5 might strictly go down moving from `NRC.v1` to `NRC.v3`, because I are removing terms, I are in fact correcting labels continuously in these revisions, which results in an improvement in recall and overall performance.

| Label Set | JOY | ANGER | FEAR | SADN. | DISG. | SURP. | # OF SENT. |
|---|---|---|---|---|---|---|---|
| (1) Ekman's labels | 0.380 | 0.125 | 0.141 | 0.226 | 0.029 | 0.098 | 1167 |
| (2) -DISGUST | 0.392 | 0.129 | 0.146 | 0.233 | - | 0.101 | 1133 |
| (3) ANGER+DISGUST | 0.369 | 0.18 | 0.137 | 0.219 | - | 0.095 | 1204 |
| (4) -SURPRISE | 0.422 | 0.139 | 0.157 | 0.251 | 0.032 | - | 1053 |
| (5) -SURPRISE, -DISGUST | 0.436 | 0.143 | 0.162 | 0.259 | - | - | 1019 |
| (6) -SURPRISE, DISGUST+ANGER | 0.407 | 0.199 | 0.151 | 0.242 | - | - | 1090 |

Table 4.7: Fairy tales label distribution

### 4.4.2 Varying the Label Sets

Alm's "high agreement" dataset only contains 148 sentences with DISGUST and SURPRISE labels, a highly imbalanced distribution. To investigate the impact of this imbalance on the results, I repeated the emotion detection experiment six times for each of the three version of the NRC, once for each of the reduced label sets shown in Table 4.6, which also shows how varying the label sets affects the performance of the emotion detection system for different version of the NRC. In all cases my corrected verisons of the NRC improve performance, anywhere from 5.3 to 7 points of $F_1$.

1. Ekman's Labels: JOY, ANGER, DISGUST, SADNESS, FEAR, SURPRISE
2. ANGER, JOY, SADNESS, FEAR, SURPRISE, dropping DISGUST-labeled sentences
3. JOY, SADNESS, FEAR, SURPRISE, combining ANGER+DISGUST
4. JOY, ANGER, DISGUST, SADNESS, FEAR, dropping SURPRISE-labeled sentences
5. ANGER, JOY, SADNESS, FEAR, dropping DISGUST- and SURPRISE-labeled sentences
6. JOY, SADNESS, FEAR, combining ANGER+DISGUST, dropping SURPRISE-labeled sentences

Table 4.7 shows the distribution of labels for Alm's data under different label sets.

Therefore, each of the 18 values in Table 4.6 is corresponding to an experiment with a different combination of emotion lexicon and label set. five labels (Joy, Anger, Disgust,

Sadness, Fear), Four labels (Anger+Disgust, Joy, Sadness, Fear), and Four labels (Anger, Joy, Sadness, Fear) and removing 34 disgust-labeled sentences from All-agreement annotated Alm's fairy tales.

## 4.5 Contributions

I noted three categories of error in the popular NRC emotion lexicon, including a large number of seemingly biased entries. I developed and applied a semi-automatic procedure to generate three different corrected version of the NRC, and showed via experiment that these new versions improved the performance of an existing emotion-lexicon-based emotion detection system. This work shows the utility of careful error checking of lexical resources, especially with attention to correcting for unintended biases. Finally, I release the revised resource and my code to enable other researchers to reproduce and build upon results[6].

[6] https://doi.org/10.34703/gzx1-9v95/PO3YGX

Chapter 5

# The ABBE Corpus: Animate Beings Being Emotional

The dataset collection process is an influential part of every machine learning work, and can affect the quality of the work significantly in both negative and positive ways. This challenge becomes more important and complicated when dealing with textual datasets because of two reasons. First, manually annotating such datasets in a large-size is a costly and time consuming process. Second, in most cases, assigning labels is related to the humans non-objective cognitive views. Emotion detection is an NLP task that has been established for quite some time, and has seen quite a few published corpora, resources, and systems [Olveres et al., 1998, Mueller, 1998, Aman and Szpakowicz, 2008, Chatterjee et al., 2019, Zad and Finlayson, 2020]. Traditionally, emotion detection consists of categorizing a piece of text as to an expressed emotion, for example, tagging the sentence "I was furious." with the label ANGER. Alternatively, the task might also involve first finding spans of text that express emotion before categorizing them, for example identifying that it is the word *furious* that provides the affective semantics for that sentence.

This approach to emotion detection is useful, but notably leaves out a key piece of information: namely, who exactly is experiencing the emotion. Emotions usually do not appear in a vacuum, and are usually experienced by *someone*, and knowing who is experiencing the emotion is a important step in understanding the semantics of the text. Accordingly I present the first corpus where emotion expressions are associated with those who are experiencing the emotion. In the general case, emotions might be associated with several different types of experiencers (e.g., the author, the narrator, the audience, etc.), each presenting their own challenges for definition and annotation. In the ABBE corpus I focus on animate beings who are part of the world of the text and are experiencing emotions, and I provide annotations on top of 30 chapters (i.e., narratives) drawn from the Corpus of English Novels, providing several thousand animate beings and expressed

emotions that can be used for training, testing, and validation of automatic systems. There are many possible reasons for doing annotation, but they all fall into two basic categories. The first reason is to capture, in an explicit and unequivocal representation, some aspect of a person's understanding of a text, so that this information may then be used to train or test computer systems that aim to emulate people. The second reason is to construct an understanding of some aspect of a text which cannot yet be automatically determined by computer. As a psychological model I chose plutchik as it is a common and basic psychological model used by researchers in the emotion detection field [Tabak and Evrim, 2016, Abdaoui et al., 2017, Rose et al., 2018, Lee et al., 2019, Ljubešić et al., 2020, Zad et al., 2021c]. Also, There is no annotated text for fairytales or narratives regarding these eight emotions. Moreover, Alm's childern fairy tales dataset is annotated based on five of Ekman's six emotions and I aimed to extend the number of emotion classes for using in future research. I reviewed in (§2) a selection of language resources that are relevant to emotion, and provide context for the ABBE corpus. I extensively reviewed both emotion lexicons and annotated corpora in (§2).

This chapter proceeds as follows. First, I review basic definitions of the key concepts in play, i.e., *emotion* and *animate being* (§5.1). Then I describe in detail the annotation scheme I designed (§5.2), as well as the texts included in the corpus, my selection criteria, and agreement measures (§5.3). I conclude with an examination of interesting and difficult edge cases (§5.4), and my contributions (§6.7).

## 5.1 Definitions

### 5.1.1 Emotion

Theories of emotion go back to the ancient Greeks and Romans, and have been a recurring theme of inquiries into the nature of the human experience throughout history, including famous proposals by Charles Darwin and William James in the 19th century [Darwin and Prodger, 1998, James, 1890]. Regardless of the spectrum of complexity stated in (§2), what is crucial for my work is that emotion is a mental state that must be attributed to a being capable of maintaining such a state.

In cognitive / psychological approaches (which, as mentioned in (§2)) there are three broad classes of theories that attempt to describe what emotions exist and the interactions between them: categorical, dimensional, and hybrid. **Categorical** models propose a discrete set of emotions; these include theories by Ekman [1999], Parrott [2001], Shaver et al. [1987], Oatley and Johnson-Laird [1987], and Izard [2007]. **Dimensional** theories propose descriptive dimensions of and relations between emotions, such that experienced or expressed emotions fall along the relevant dimensions and potentially shade into each other, and are not necessarily distinct. Theories in this class include those by Russell and Barrett [1999], Scherer [2005], Lövheim [2012], Ortony et al. [1990], and Fontaine et al. [2007]. Finally, there are **Hybrid** models which combine aspects of both categorical and dimensional theories; the theory by Plutchik [Plutchik, 2001a] falls into this class. In my work, for reasons given in section below, I adopt the eight basic categories of Plutchik's model. Regardless of the specific theory chosen, however, what remains is the importance of the concept of mental state, which means that a proper description of emotion includes not only the emotion itself, but also the experiencer of an emotion. I turn to that next.

### Plutchik's Wheel of Emotions Model

As mentioned above, there are numerous psychological theories of emotion. In this work I have chosen to use the eight basic emotions of Plutchik's model. The reasons for this choice are twofold. First, the most commonly used emotion theory in NLP is Ekman's six category model [Ekman, 1999]; however, this model has several noted deficiencies including lacking several key emotional concepts (i.e., trust, anticipation) as well as lacking any well defined relationship between the categories. Plutchik's model resolves these two problems without adding a significant amount of complexity (which would increase the difficult of annotation), while still being compatible with Ekman's model. Second, there are a number of available resources for Plutchik's model, including both emotion lexicons and corpora, which make the choice of this as a model practical in terms of building immediately useful systems.

To understand Plutchik's model, shown in Figure 2.3, it is necessary to break it down based on the various aspects of the model, such as what the primary emotions are and their opposites.

**Primary** The eight colored sections are designed to indicate that there are eight primary emotions: anger, anticipation, joy, trust, fear, surprise, sadness and disgust. Some attributes that are associated with these eight sectors can be seen in Table 5.1. The eight primary emotions can be seen related to certain scenarios and cognition that help others understand what emotion an individual is experiencing in each moment and possible reasons for it. A person who feels as if their life is threatened, as shown in Table 5.1, would most likely feel fear of dying and try to leave the situation to feel safe again. Throughout this, normally a person would feel fear and would express such emotion. This is further supported in Table 5.2 as shown with certain cases such as reacting to contact with a strange object can surprise people. This can be seen too with anger when someone is de-

| Stimulus Event | Cognition | Feeling State | Overt Behavior | Effect |
|---|---|---|---|---|
| Threat | Danger | Fear | Escape | Safety |
| Obstacle | Enemy | Anger | Attack | Destroy Obstacle |
| Gain of valued object | Possess | Joy | Retain or repeat | Gain resources |
| Loss of valued object | Abandonment | Sadness | Cry | Reattach to lost object |
| Member of one's group | Friend | Acceptance (Trust) | Groom | Mutual support |
| Unpalatable object | Poison | Disgust | Vomit | Eject poison |
| New territory | Examine | Expectation (Anticipation) | Map | Knowledge of territory |
| Unexpected event | What is it? | Surprise | Stop | Gain time to orient |

Table 5.1: Related attributes of feeling states [Plutchik, 2001b].

structive. Emotions can be shown through the actions, words, responses, and more from an individual in a certain situation.

**Opposite** Each primary emotion has a polar opposite:

- Joy $\longleftrightarrow$ Sadness
- Fear $\longleftrightarrow$ Anger
- Anticipation $\longleftrightarrow$ Surprise
- Disgust $\longleftrightarrow$ Trust

**Combinations** Emotions are often complex. The emotions placed between the colored sections are those represented as a mix of the two neighboring primary emotions. For example, Anticipation and Joy combine into Optimism. Joy and Trust combine into Love.

**Intensity** The vertical dimension, shown radially in the main portion of the figure, and vertically in the upper left inset, Figure 2.3, represents intensity—emotions intensify as they move from the outside to the center of the wheel (top to bottom of the cone), which is also indicated by the color: The darker the shade, the more intense the emotion. For example, Anger at its lowest level of intensity is Annoyance. At its highest level of intensity, Anger becomes rage. Another is a feeling of Boredom, which can intensify to loathing if left unchecked, noted as dark purple.

| Prototype Adaptation | Hypothesized Emotion |
|---|---|
| Protection: Withdrawal, retreat, contraction | Fear, Terror |
| Destruction: Elimination of barriers to the satisfaction of needs | Anger, Rage |
| Incorporation: Ingesting nourishment | Acceptance (Trust) |
| Rejection: Riddance response to harmful material | Disgust |
| Reproduction: Approach, contact, genetic exchanges | Joy, Pleasure |
| Reintegration: Reaction to loss of a nutrient object | Sadness, Grief |
| Exploration: Investigation of one's environment | Curiosity, Play (Anticipation) |
| Orientation: Reaction to contact with a strange object | Surprise |

Table 5.2: Hypothesized Emotions [Plutchik, 2001b].

### 5.1.2 Experiencers of an Emotion: Animate Beings

Key to emotion is mental state, and to have a mental state there must be a mind. There are numerous minds—real or imagined—that could be the experiencer of emotions described in text. From the point of view of a reader of a text, possibly the first being that comes to mind as potentially experiencing an emotion is the **reader** themself: individuals who consume a text, be it a narrative, essay, textbook, or other genre of text, can experience emotions during that consumption. These emotional experiences may or may not correspond to emotions described in the text—for example, a student might feel despair or anguish on reading the first paragraphs of their new textbook on statistical mechanics, despite those emotions not being directly describe in the textbook in question. There, of course, are notable cases where emotions described in a text might reasonably be expected to be experienced by the reader, such as the case in literature when readers feel sympathetic emotion that mirrors that being experience by characters in the narrative. This is obviously a difficult case because how an individual or an "average reader" (if such a person can even be reasonably constructed) might react to a particular text can be extremely difficult to predict.

Another possible emotion experiencer is the **author** (as distinct from the presented narrator) of the text. In these cases, especially in cases of first person description, the

author may describe or imply emotions attributable to themself. This can be communicated explicitly through semantically emotional words (e.g., *When writing this book, I fell into depression.*) or may be expressed implicitly through devices such as style or even punctuation (e.g., the use of explanation points).

A third broad category of experiencer I will term **animate beings**, by which I specifically mean beings described as being part of the "world" of the text. Canonically, such beings are often thought of as the "characters" in the world of the text; however, my category is broader than character because there might beings described in the text which can experience emotions which are not, narratively speaking, characters. I thus more precisely define this concept.

I start by defining *animacy*, which is the characteristic of being able to independently carry out actions (e.g., movement, communication, etc.) [Jahan et al., 2018b]. Human beings, for example, are animate because they can move and communicate in a realistic environment; however, a chair or a table cannot do these things on their own, hence they are typically regarded inanimate. Animacy is a required property of characters in stories, which means that all characters must be animate in the traditional sense.

Characters, however, are not the only possible animate beings in a text. As defined in detail elsewhere [Jahan et al., 2020b], characters are *animate beings that are important to the plot of a narrative*, meaning that they have a non-trivial role in advancing the action described in the text. With this distinction in mind, one can see that not all beings mentioned in a text are necessarily character: the text in question might not actually be a proper narrative, or there might be other, minor beings that could in theory be removed while keeping the essence of the plot or action of the text. I include these beings in my definition of *animate being*. An animate being, then, is any entity described in a text that can act autonomously or individually such as a person, an animal, the narrator, an imaginary creature, a magically animate tree, etc. Any and all of these beings can potentially experience emotions in the text.

The challenges of identifying the emotions experienced by these three classes of potential experiencers of emotion (readers, authors, and animate beings) differs greatly; therefore, I concentrate in this work on the third class (animate beings), as these seem to account for the vast majority of explicit references to emotion in texts commonly encountered.

## 5.2 Annotation Scheme & Process

I detail here the different parts and definitions of the ABBE annotation scheme, and describe how the corpus was annotated.

### 5.2.1 Annotation Scheme

There are four components to annotation scheme: (1) the emotional expression span; (2) the emotion expressed in that span; (3) animate beings experiencing the identified emotions.

#### Emotional Expression Span

Annotators were first asked to identify spans of text that expressed emotions. They read the text and identified any emotional words or phrases. They were asked to identify the minimum span of contiguous tokens that covered all the emotional words in a single expression. Annotaters were allowed to reference WordNet [Fellbaum, 1998b] to help them disambiguate when words were being used in an emotional sense.

(1) "John was the [happiest] man alive": In this example, I mark only the token *happiest* as emotional, not including the article or the modified phrase *man alive*, as

these are outside the minimum span of contiguous tokens that cover all the emotional words.

(2) In contrast, "Edna's death has filled me with immense [sorrow and dread] as I saw her life-less corpse": The emotional span covers *sorrow and dread* since multiple tokens, separated by non-emotion words (i.e., *and*), are used to highlight the emotion occurring in this particular instance of the text which pertains to an article.

**Emotion Expressed**

Once the span was identified, annotators were asked to mark the span as to the emotion or emotions expressed, making this a multi-label classification task. As discussed, I used Plutchik's primary taxonomy of eight emotions: Joy, Trust, Fear, Surprise, Sadness, Anticipation, Anger, and Disgust. I provide descriptions of each category below.

**Joy** A feeling of extreme gladness delight, or exaltation of the spirit arising from a sense of well-being.

(3) "The young man was fluent and [gay]$^{\text{Joy}}$, but he laughed louder than was natural in a person of polite breeding…"

**Sadness** An emotional state of unhappiness usually aroused by the loss of something that is highly valued

(4) "It was only forty pounds he needed," said the young man [gloomily]$^{\text{Sadness}}$.

**Disgust** A strong aversion to something deemed revolting, or towards a person's behavior deemed repugnant

(5) From the whole tone of the young man's statement it was plain that he harbored very [bitter and contemptuous]$^{\text{Disgust}}$ thoughts about himself.

**Trust** A reliance on or confidence in the dependability of something.

(6) I [confide]$^{\text{Trust}}$ in your abilities to get through all of this chaos.

**Anger** An emotion characterized by tension and hostility arising from frustration.

(7) All you do is [anger]$^{\text{Anger}}$ me everyday.

**Fear** An intense emotion aroused by the detection of imminent threat, involving an immediate alarm reaction.

(8) "Your Highness," said the Colonel, turning [pale]$^{\text{Fear}}$; "let me ask you to consider the importance of your life, not only to your friends, but to the public interest."

**Anticipation** A looking forward to a future event or state, with an affective component.

(9) He was eagerly waiting for it, [expecting]$^{\text{Anticipation}}$ it to come for him.

**Surprise** An emotion resulting from the violation of an expectation or the detection of novelty.

(10) At its close Lady Theobald found herself in an utterly [bewildered and thunder-struck]$^{\text{Surprise}}$ condition.

As mentioned, annotators were allowed to assign multiple emotion categories if a span expressed more than one emotion.

(11) The poor man was [jealous]$^{\text{Anger, Disgust}}$ of the rich kid winning the lottery.

(12) Remembering the tragic moment has instilled his soul with [anger and sorrow]$^{\text{Anger, Sadness}}$ as he realized how corrupt the world around him was.

(13) He could not contain himself, as he ran towards the entrance of the park [gleefully looking forward]$^{\text{Joy, Anticipation}}$ to the adventures today will bring.

**Identifying the Emotion Experiencer**

With labeled emotional expression spans in hand, the last part of the annotation is to identify the animate being who is experiencing the emotion. To identify animate beings, I followed the same procedure as specified by Jahan et al. [2018b]. In particular, as I based ABBE off of the 30 selections from the Corpus of English Novels already annotated in that work so that I could check our annotations. The annotators looked for the closest referring expression for the relevant animate being based on their understanding of the text. If no animate being could be identified, the emotional span was dropped from the dataset. I found only minor variations between our annotations of animate beings and those of Jahan et al.

In the end, this sequence of steps resulted in annotations structured as follows:

(14) [His]$^{\text{AB1}}$ thoughts were both quiet and [happy]]$^{\text{Joy}\rightarrow\text{AB1}}$ His brief favour with the Duke he could not find it in his heart to mourn; with Joan to wife, and my Lord Foxham for a faithful patron, [he]$^{\text{AB2}}$ looked most [happily]$^{\text{Joy}\rightarrow\text{AB2}}$ upon the future; and in the past he found but little to regret.

In this short snippet of text there are two emotional experiences, the first being experienced by the person referred to by *His*, and the second by *he*. It just so happens that

these two referring expressions are co-referent; I did not mark coreference explicitly in the dataset, but it can be extracted from Jahan et al.'s annotations of the same texts.

### 5.2.2 Annotation Workflow

As mentioned previously, the ABBE corpus is double-annotated. Two annotators (the first and second authors) performed the annotation. Each week, both annotators were given the same, specific collection of text to annotate. Once both annotators were finished, they met to review and adjudicate disagreements.

### 5.2.3 Agreement Measures

Agreement on identification of the tokens that are part of emotion spans, before adjudication, was 0.933 $F_1$. Agreement on identification of animate beings, before adjudication, was 0.970 $F_1$. These two sets of judgements were then adjudicated, which allowed us to compute inter-annotator agreement for emotion assignment.

The overall inter-annotator agreement on emotion assignment was 0.826 measured using Cohen's kappa ($\kappa$). Cohen's kappa measures the agreement between two raters who each classify $N$ items into $C$ categories; here, there are $C = 8$ different emotion categories. $\kappa$ is defined as $\frac{p_o - p_e}{1 - p_e}$ where $p_o$ is the relative observed agreement. Assuming that emotion sets $L_1$ and $L_2$ represent the multi-emotion labels given by raters 1 and 2 respectively, the relative observed agreement is calculated as:

$$p_o = \frac{|(L_1 \cap L_2) \cup (\overline{L_1} \cap \overline{L_2})|}{8} \tag{5.1}$$

where 8 is the number of emotion labels. $p_e$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. I used the following formula to calculate $p_e$:

$$p_e = \frac{\sum_k N_{1k} N_{2k}}{N^2} \tag{5.2}$$

where $N_{1k}$ and $N_{2k}$ are the number of times that raters 1 and 2 predicted category $k$.

## 5.3 Selected Texts

ABBE comprises 30 selections from the Corpus of English Novels, a 25-million-word corpus which consists of 292 novels by 25 different novelists [De Smet, 2008a]. These novels were written between 1881 and 1922 and the corpus was designed to track short-term language changes and comparing usage across a certain generation of authors. The purpose of selecting novels for ABBE is to take advantage of the relative density and variety of reported emotion due to the nature of the genre and a large number of different characters, environments, situations, and writing styles provided by the different novels from different writers. The key counts for the corpus are given in Table 5.3, while the specific selections included in ABBE are shown in Table 5.4

| **Feature** | **Count** |
|---|---|
| # of Texts | 30 |
| # Tokens | 134,513 |
| # Emotional Animate Beings | 2,227 |
| # Emotion Spans | 2,010 |
| # Joy labels | 735 |
| # Sadness labels | 678 |
| # Disgust labels | 359 |
| # Trust labels | 471 |
| # Anger labels | 408 |
| # Fear labels | 540 |
| # Anticipation labels | 649 |
| # Surprise labels | 409 |

Table 5.3: Key Counts for the ABBE Corpus

| Novels | Name Of Story | Author | Chapters |
|---|---|---|---|
| 1 | A Fair Barbarian | Frances Hodgson Burnett | 25-26 |
| 2 | Milly and Olly | Mary Augusta Ward | 10 |
| 3 | Treasure Island | Robert Louis Stevenson | 11 |
| 4 | Mr. Isaacs | Francis Marion Crawford | 14 |
| 5 | The Suicide Club | Robert Louis Stevenson | 1 |
| 6 | Doctor Claudius: A True Story | Francis Marion Crawford | 20 |
| 7 | A Roman Singer | Francis Marion Crawford | 24 |
| 8 | Miss Bretherton | Mary Augusta Ward | 7 |
| 9 | Philistia | Charles Grant Blairfindie Allen | 2 |
| 10 | The Black Arrow: A Tale of the Two Roses | Robert Louis Stevenson | 7 |
| 11 | The Unclassed | George Robert Gissing | 38 |
| 12 | A Mummer's Wife | George Augustus Moore | 30 |
| 13 | King Solomon's Mines | Henry Rider Haggard | 20 |
| 14 | Little Lord Fauntleroy | Frances Hodgson Burnett | 1 |
| 15 | Prince Otto | Robert Louis Stevenson | 4 |
| 16 | The Children of the King | Francis Marion Crawford | 12 |
| 17 | The Shadow of a Crime | Thomas Henry Hall Caine | 51 |
| 18 | Zoroaster | Francis Marion Crawford | 20 |
| 19 | A Tale of a Lonely Parish | Francis Marion Crawford | 24 |
| 20 | Demos | George Robert Gissing | 36 |
| 21 | On Being in Love | Jerome Klapka Jerome | 2 |
| 22 | Kidnapped | Robert Louis Stevenson | 30 |
| 23 | Muslin | George Augustus Moore | 29 |
| 24 | A Romance Of Two Worlds | Marie Corelli | 14 |
| 25 | Strange Case of Dr. Jekyll and Mr. Hyde | Robert Louis Stevenson | 1 |
| 26 | Vendetta!: Or The Story of One Forgotten, a Novel, Volume 1 | Marie Corelli | 1 |
| 27 | A Mere Accident | George Augustus Moore | 9 |
| 28 | Marzio's Crucifix, Volume 1 | Francis Marion Crawford | 11 |
| 29 | A Little Princess | Frances Hodgson Burnett | 18 |
| 30 | Saracinesca | Francis Marion Crawford | 34 |

Table 5.4: CEN 30 Selected Texts

## 5.4 Difficult and Interesting Cases

### 5.4.1 Multiple Conflicting Emotions

Naturally it is possible for an animate being to feel different, even conflicting, emotions over time. There is no problem when these emotion mentions are separated in the text, but in certain cases the text presents these emotions as being attached to the same referring expression. In these cases I do not provide any distinguishing information, and just annotation the emotion mentions normally:

(15) [Henry]$^{AB1}$ was [upset]$^{Sadness \rightarrow AB1}$ at his grade for his midterm, but this turned to [joy]$^{Joy \rightarrow AB1}$ when it turned out to actually be a passing grade.

(16) [Anger]$^{Anger \rightarrow AB1}$ filled [John's]$^{AB1}$ veins, which soon became tears of [sorrow]$^{Sadness \rightarrow AB1}$ ...

In both examples, the relevant animate being moves from one emotion to another, but the emotions are attached to the same referring expression.

### 5.4.2 Sets of Animate Beings

It is not uncommon for a single emotional span to be attributable to several animate beings at once. In the case of separable mentions, the annotation points to each individual animate being. For example:

(17) [Jack]$^{AB1}$ was afraid of [falling]$^{Fear \rightarrow AB1,AB2}$, and so was [Jill]$^{AB2}$.

In cases where the animate beings are indicated by a single phrase (such as a conjunctive noun phrase, or a plural pronoun), the phrase is marked as the relevant animate being:

(18) [Jack and Jill]$^{AB1}$ were [scared]$^{Fear \rightarrow AB1}$ of falling.

(19) [They]$^{AB1}$ were [scared]$^{Fear \rightarrow AB1}$ of falling.

In these cases, the emotion is attributed to the set that comprises *Jack and Jill* (or *they*).

### 5.4.3 Emotion vs. Action

There is a difference between emotion and action; while there are many actions which imply the agent is or will experience certain emotions, I don't mark an emotion unless it

is necessarily part of the semantics of the action, or the emotion is explicitly mentioned. Compare the following:

(20) I went to Disneyland. (no emotion)

(21) [I]$^{AB1}$ was [ecstatic]$^{Joy \rightarrow AB1}$ to go to Disneyland.

### 5.4.4 Emotion vs. Mood

There are also cases where an emotion is not present but is generalized, not isolated in time, and cannot be assigned to a specific animate being or set of animate beings. These I call *moods*, and I do not mark them. For example:

(22) Dread fills the city late at night.

In this sentence, Dread is clearly affective, but there is no one specifically mentioned as experiencing the emotion.

## 5.5 Contributions

My contributions in this chapter are threefold. First, I note the importance of the experiencer to the idea of emotion, and point out that all prior corpora of annotation emotion omit this information. Second, I define an annotation scheme that captures the experiencer of an emotion (an animate being), based on prior work on emotion classification and animate being detection. Finally, I provide the ABBE corpus, a collection of 30 texts (134.5k tokens) annotated for 2,010 emotion spans associated with 2,227 animate beings, annotated with excellent inter-annotator agreement.

Chapter 6

# Detecting the Emotions of Animate Beings

Emotion is an important feature of communication, especially so for narratives [Herman, 2014]. Importantly, animate beings are often the vehicles of emotional expression in narrative: animate beings, especially characters, are key elements of narrative [Chatman, 1986, Margolin, 1990] and are often portrayed as experiencing particular emotions, which are separate from—or intended to stimulate—emotions in either the narrator figure or the reader. Therefore, detecting the emotions experienced by animate beings in a narrative is an important step toward general automatic narrative text understanding.

Emotion detection in text has received quite a bit of attention, leading to a number of approaches for automatically detecting emotion expressed in text [Binali et al., 2010, Shelke, 2014, Canales and Martínez-Barco, 2014, Garcia-Garcia et al., 2017, Sailunaz et al., 2018, Zad et al., 2021b], including narrative text specifically [Alm, 2008, Zad and Finlayson, 2020]. Similarly, there exist recently developed approaches to detecting animate beings [Jahan et al., 2018a, 2020a]. However, no one has yet combined these approaches to detect emotions represented as being experienced by the animate beings in a story. I present such a system here.

The canonical case of such an expression might be something like "Emma Bovary felt happy that day.", where the animate being is *Emma Bovary* and the emotion felt by that being is *joy* (in Plutchik's ontology). The simplicity of this canonical example belies the difficulty of the task: my system achieves an overall $F_1$ of 0.76, which, while respectable, clearly indicates there is room for improvement.

I begin this chapter to describe the dataset I use for training and testing (§6.1) and then describe my emotion detection pipeline (§6.2), which leads to a presentation of the results (§6.5). Finally, I identify some unsolved challenges that point toward future work (§6.6) and summarize my contributions (§6.7).

## 6.1 Dataset: ABBE

For training and testing I used the recently released Animate Beings Being Emotional or Emotional Animate beings (ABBE) dataset [Zad et al., 2022]. This dataset contains 30 chapters, each a chapter of a book drawn from the Corpus of English Novels [CEN; De Smet, 2008b], comprising a total of 134,513 words. The base dataset was created by Jahan et al. [2018a] and was annotated for coreference structure as well as animacy, allowing the identification of animate beings. ABBE then adds the identification of spans of text that represent emotions experienced by animate beings described in the text, marks them with one or more emotion classes drawn from Plutchik's 8-class scheme, and links those labeled spans with the appropriate animate being. ABBE contains 12,686 chains of animate beings identified by Jahan et al. [2018a]. There are 2,010 unique emotion spans identified, with an average of 1.85 emotion labels assigned to each span. Zad et al. reported inter-rater reliability scores of 93.5 $F_1$ for identifying emotional spans, $\kappa = 88.08\%$ for assigning those spans labels, and 0.988 $F_1$ for associated spans with beings.

## 6.2 Overview of the Emotion Detection System

My system combines multiple modules for pre-processing, semantic role labeling, animate being detection, and emotion labeling to achieve detection of the emotions described as being experienced by animate beings in the text. The fully automatic sequence of steps is as follows.

1. **Pre-Processing**: Using the Stanford CoreNLP library [Manning et al., 2014] I pre-processed the 30 chapters in the corpus by identifying sentence boundaries, tok-

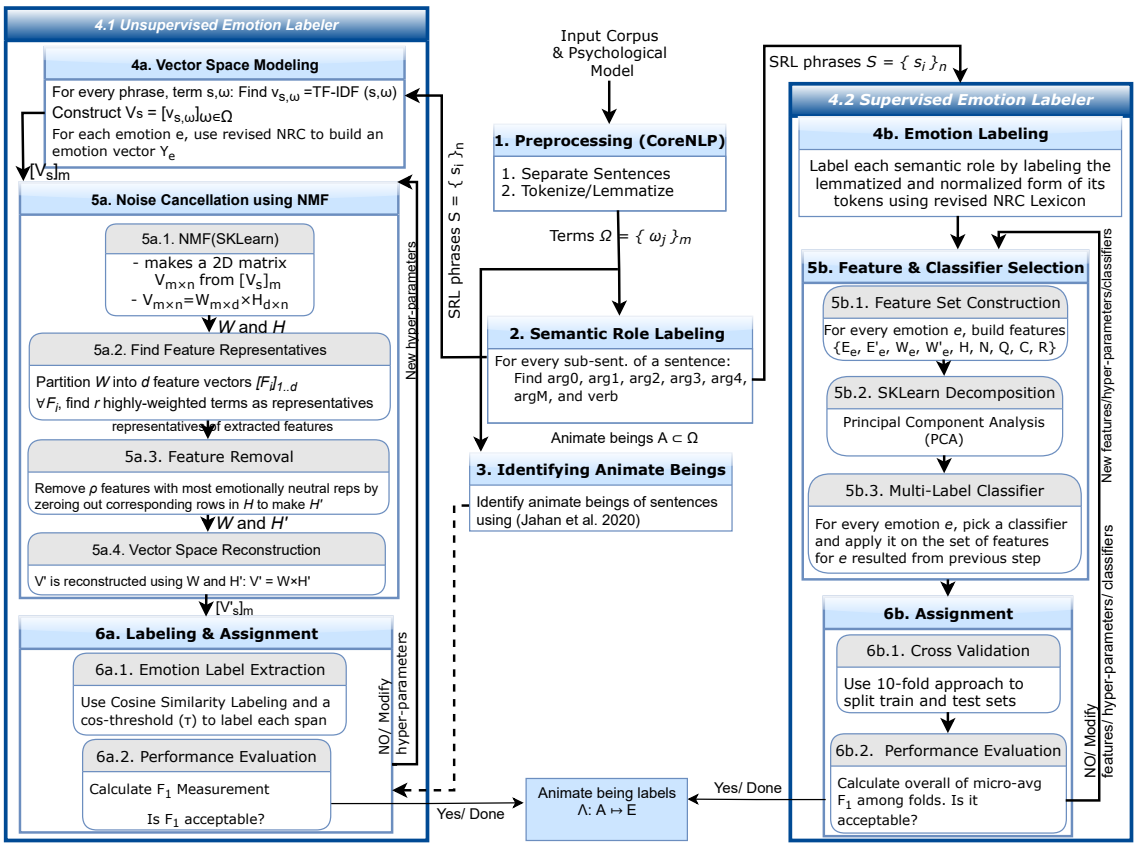



Figure 6.1: The automatic procedure for animate being's emotion detection.

enizing each sentence, computing lemmas for each token, and pos-tagging each lemma.

2. **Semantic Role Labeling**: I applied semantic role labeling (SRL) to find the verbal arguments and their roles in sentences of the corpus using the tool developed by [Shi and Lin, 2019], which achieves a $F_1$ of 78.9% on English Ontonotes 5.0 dataset.
3. **Animacy Labeling**: I labeled the animacy of coreference chains occurred in the 30 chapters using the classifier described by Jahan et al. [2020a].

### 6.2.1 Unsupervised Emotion Labeler

In steps 4 to 6, I explored two different approaches to performing emotion labeling for the animate-beings referenced in the text: (a) unsupervised, and (b) supervised. I give an overview first of the unsupervised approach; details for the below steps are given in Section 6.3.

4a. **Vector Space Modeling**: Using the POS-tagged lemmas obtained in the first step, I considered each identified SRL argument as a bag of words (BoW) and calculated a vector of tf-idf values for it. Also, I used the Revised NRC lexicon from Zad et al. [2021c] to build a set of emotion vectors for eight classes derived from Plutchick's scheme.

5a. **Noise Cancellation using NMF**: Using non-negative matrix factorization, I removed the non-affective features of my constructed vector space model to enhance the accuracy of labeling made in the next step. This follows the approach described in [Zad and Finlayson, 2020].

6a. **Labeling and Assignment**: Next I used the assignment approach described in Zad and Finlayson [2020] (with 8 emotions instead of 4) to assign a set of emotion labels to each of the arguments identified by the SRL using a cosine-similarity metric. Arguments inside the same verbal semantic role structure as an animate being were then assigned to that animate being.

### 6.2.2 Supervised Emotion Labeler

I also explored a supervised approach, outlined as follows.; Details for the below steps are given in Section 6.4.

4b. **Emotion Labeling** I used the Revised NRC lexicon from Zad et al. [2021c] to assign a set of emotion labels to each of the arguments identified by the SRL.

5b. **Feature & Classifier Selection** I used a simple, manual boosting approach to find a high-performing feature set and classifier for each label, where the classifier possibilities were drawn from the set {MLP, Decision Tree, SVM, Naive Bayes, Random Forests, and KNN}) with a variety of possible hyper-parameter settings. The classifier selected for each label was trained in a one-vs-many setup, and labels assigned to a span were combined to allow for multi-label results.

6b. **Assignment** Emotion spans inside the same verbal semantic role structure as an animate being were then assigned to that animate being.

Table 6.1: List of classifiers and their tunable hyper-parameters.

| Classifier | Hyper-Parameters |
|---|---|
| MLP | hidden layers={$(100), (50, n, 50)_{n=25}^{125}$}<br>learning rate={constant, adaptive}<br>activation={relu, tanh, logistic}<br>solver={adam, sgd, lbfgs} |
| Decision Tree | splitter={random, best}<br>criterion={gini, entropy}<br>max-depth={$3, 4, \ldots$} |
| SVM | kernel={linear, poly, rbf, sigmoid}<br>gamma={scale, auto} |
| Gaussian NB | var_smoothing={10e-$n$}$_{n=5}^{12}$ |
| Rand. Forest | n_estimators={$50, 60, \ldots, 150$}<br>criterion={gini, entropy} |
| KNN | n_neighbors={3, 4, ..., 15}<br>weight={uniform, distance}<br>alg.={auto, ball-tree, kd-tree} |

## 6.3 Details of Unsupervised Emotion Labeler

### Step 4a: Vector Space Modeling of Semantic Roles

For every phrase $p$ labeled as a semantic argument by the SRL, I construct a bag of `(lemma, pos-tag)` pairs by tokenizing, lemmatizing, and POS tagging each word of

the phrase. Assuming that $\Omega$ represents the set of all `(lemma, pos-tag)` pairs generated in this process across all phrases, I construct a count vector for $\text{BoW}_p$ by mapping each lemma, pos-tag to the count in the phrase ($\Omega \mapsto \mathbb{Z}_{\geq 0}$).

Using the count vectors so constructed, I compute a *tf-idf* vector for each phrase. For each phrase $p$, I construct an $m$ dimensional vector where each entry in the vector is the *tf-idf* of `(lemma, pos-tag)` pair $\omega_i$ in phrase $p_j$; i.e.

$$v_{ij} = -\text{BoW}_{p_j}(\omega_i) \times \log \frac{\text{pop}(\omega_i)}{n} \tag{6.1}$$

where $n$ is the number of all phrases ($|P|$), $\Omega = \{\omega_i\}_{i=1}^{m}$, and

$$\text{pop}(\omega_i) = |\{p \in P : BoW_p(\omega_i) > 0\}|. \tag{6.2}$$

The constructed vector space model is represented by an $m \times n$ matrix $V = [V_{p_j}]_{j=1}^{n}$ where $V_{p_j} = (v_{1j}, v_{2j}, \ldots, v_{mj})^T$.

I also compute a "standard vector" for each of Plutchik's eight emotion classes $Y_e = (y_{e,\omega_1}, y_{e,\omega_2}, \ldots, y_{e,\omega_m})^T$ where $y_{e,\omega_i}$ is 1 if the (lemma, pos-tag) pair $\omega_i$ is mapped to $e$ by the lexicon, otherwise 0.

## Step 5a: Noise Cancellation using NMF

The vectors $V_p$ and $Y_e$ from the previous step are all $m$-dimensional vectors where $m$ is the total number of terms in the corpus. There are many terms that have little or no effect on the emotion labeling of the animate-beings in their sentences. Therefore, the noise cancellation technique Non-Negative Matrix Factorization (NMF) used by Zad and Finlayson [2020] can enhance the accuracy of emotion labeling system.

When using NMF for decomposing the vector space model, $V$ is factorized into two matrices $W_{m \times d} = [w_{ij}]$ and $H_{d \times n} = [h_{ij}]$, both with all non-negative entries. In this

decomposition, integer $d$ is considered a hyper-parameter and its numerical value can be fine-tuned by maximizing the accuracy of the labeling system on a certain data set.

The NMF factorization process produces a matrix $W$ whose $d$ columns each represents an $m$-dimensional feature for each of the $n$ phrases in the corpus. In other words, $W = [F_j]_{j=1}^{d}$ where $F_j = (w_{1j}, w_{2j}, \ldots, w_{mj})^T$ for each $j = 1, 2, \ldots, d$, where each of the $d$ rows of $H$ matrix represents weights of the $d$ features in $F$.

For every feature $F_j$, I identify a fraction $r$ of $\omega_i$ pairs with the highest weights as its representatives, where $r$ is a hyper-parameter that can be tuned during system optimization ($r$ is usually less than 0.01).

In the next step, I remove the $\rho$ features that have little or no emotional relevance, where $\rho$ is a non-negative integer hyper-parameter that can be tuned. I will call a feature "emotionally irrelevant" if all of its representative terms are labeled as neutral by the lexicon. These features will always be removed first. If $\rho$ is less than the number of emotionally irrelevant features, I choose at random. On the other hand, if the number of emotionally irrelevant features is less than $\rho$, I eliminate features $F_j$ in order of their overall emotional relevance, which is computed by estimating the standard deviation of values $\{\eta_{je} \mid e \in E\}$ where $\eta_{je}$ is number of $F_j$ representatives labeled by $e$ in the revised NRC and $E$ is the set of eight emotion classes defined in Plutchik's wheel.

Next, the vector space model is reconstructed ($V'$) after eliminating the irrelevant features. Let $I$ denote the set of indices whose corresponding features are identified as least relevant in previous step ($|I| = \rho$). Then, the reconstructed vector space is:

$$V' = [v'_{ij}]_{m \times n} \text{ s.t. } v'_{ij} = \sum_{1 \leq k \leq d,\ k \notin I} w_{ik} h_{kj} \tag{6.3}$$

### Step 6a: Labeling & Assignment

Emotion labeling is effected by measuring the similarity between updated vectors of phrases $[V'_p]_{p \in P}$ and standard emotion vectors $[Y_e]_{e \in E}$. Assuming that $\tau$ is a hyper-parameter referred to as the *cosine-similarity threshold* (and which can be tuned like other parameters), the multi-emotion label of each phrase $p$ is calculated using the following formula:

$$\text{predicted labels of } p = \{e \mid \text{sim}(V_p, Y_e) \geq \tau\} \tag{6.4}$$

where similarity function can be measured by the cosine of angle made by the two given vectors and the value of $\tau$ is generally between 0.01 and 0.1. I then using the semantic role structures to directly assign emotion spans to animate beings.

To enhance the performance of the system, I tuned the parameters listed in steps 5a $(d, r, \rho)$ and 6a $(\tau)$ using 10-fold cross-validation.

## 6.4 Details of Supervised Emotion Labeler

Although the unsupervised approach works well, I explored the use of supervised models to achieve the same result.

### Step 4b: Emotion Labeling of Semantic Roles

To generate emotion labels for semantic role arguments (generated in Step 2) to use for supervised learning, I used the revised NRC lexicon [Zad et al., 2021c] which assigns every `(lemma, POS-tag)` pair $\omega_i$ to a subset of eight emotions defined in Plutchik's wheel [Plutchik, 1984]. Given the set of all $\omega_i$ in a semantic role $R$, the emotion label $e$ is assigned to $R$ if and only if one $\omega_i$ in $R$ is labeled by $e$ in the Revised NRC. Therefore, a semantic role can get zero, one, or multiple emotion labels in this step.

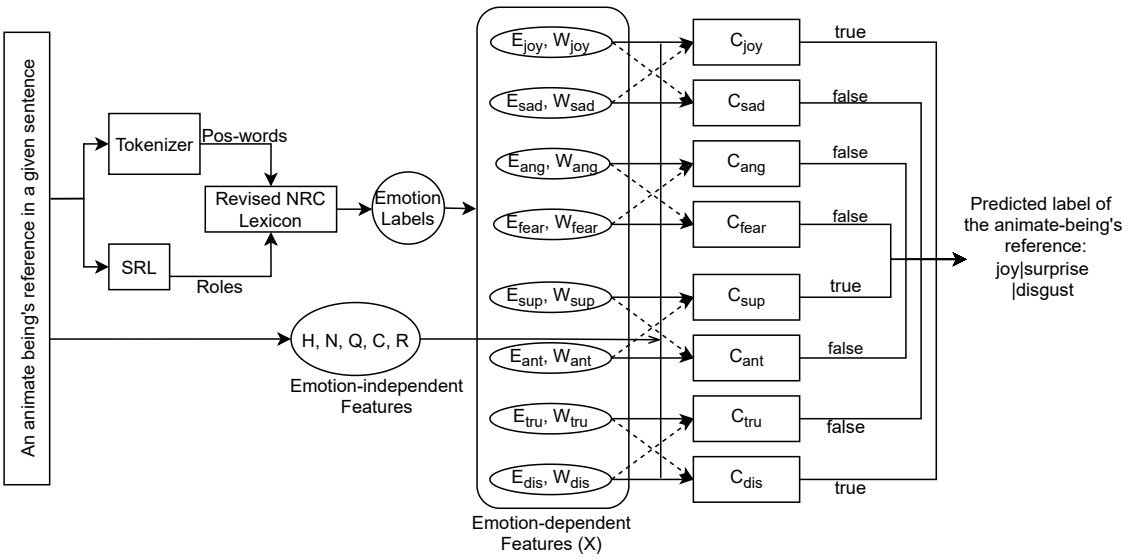



Figure 6.2: Details of multi-label emotion classifier .

## Step 5b: Feature & Classifier Selection

### Step 5b.1: Feature Set Construction

I use the outcome of steps 1 and 5a to build, extract and engineer the features of a binary classifier represented by $C_e(\mathcal{E}_e, \mathcal{W}_e, \mathcal{H}, \mathcal{N}, \mathcal{Q}, \mathcal{C}, \mathcal{R}, \mathcal{X})$ for every emotion label $e$ in Plutchik's emotion model set containing joy, anger, disgust, sadness, fear, anticipation, trust, and surprise. The outcome of classifier $C_e$ determines whether a given animate being's reference is labeled by emotion $e$ (true) or not (false) as illustrated in Figure 6.2. Here is the description of every possible feature in the feature set:

- $\mathcal{E}_e$: a vector of binary (true/false) values $b_0, b_1, \ldots, b_5, b_m, b_v$. For each $i = 0, 1, \ldots, 5$, $b_i$ determines whether the animate being is part of a semantic role labeling such that role $\text{ARG}_i$ is labeled by emotion $e$. Also, $b_m$ and $b_v$ are similar values specifying the emotion label of ARGM and the verb.
- $\mathcal{W}_e$: a binary value determining whether there exists at least one emotional word labeled by $e$ in the sentence containing the reference to animate being.

- $\mathcal{H}$: a binary value specifying whether the reference to animate being is at the beginning (head) of the sentence.
- $\mathcal{N}$: two binary values, one determines whether the reference to animate being is associated to a negated phrase by my semantic role labeler, and the second one specifies whether the reference is part of a sentence containing a negated phrase.
- $\mathcal{Q}$: a binary value determining whether the reference to animate being is part of a question.
- $\mathcal{C}$: two binary values, one determines whether the reference to animate being is associated to a conditional statement by my semantic role labeler and the second one specifies whether the reference is part of a sentence containing a conditional statement.
- $\mathcal{R}$: semantic role of the reference to animate being determined by my semantic role labeler. The role can be ARG0, ARG1, ..., ARG5, or ARGM.
- $\mathcal{X}$: In Plutchik's Wheel, the eight emotion labels are grouped into four pairs of opposite labels: (joy, sadness), (anger, fear), (surprise, anticipation), and (trust, disgust). For every emotion label $e$, let $\bar{e}$ represent the opposite emotion of $e$. I define the two features $\mathcal{E}_{\bar{e}}$ and $\mathcal{W}_{\bar{e}}$ for the classifier $C_e$ in a similar way that $\mathcal{E}_e$ and $\mathcal{W}_e$. These features determine whether the targeted sentence contains a word/phrase labeled with opposite emotion; e.g. $C_{\text{joy}}$ can benefit from vectors $\mathcal{E}_{\text{sadness}}$ and $\mathcal{W}_{\text{sadness}}$. In order to take advantage of these two features for any given emotion, I *may* replace $\mathcal{E}_e$ and $\mathcal{W}_e$ by $\mathcal{E}'_e = \mathcal{E}_e - \mathcal{E}_{\bar{e}}$ and $\mathcal{W}'_e = \mathcal{W}_e - \mathcal{W}_{\bar{e}}$ respectively.

**Step 5b.2: SKLearn Decomposition**

I used Principal Component Analysis (PCA) [Jolliffe, 2005] to reduce the dimension of the data while retaining most of the variation in the data set. I apply PCA on the con-

structed features with the number of components chosen as part of the brute-force search to find the best performance of the multi-label classification task.

## Step 5b.3: Multi-Label Classifier

For every emotion $e$, I pick a classifier listed in Table 6.1 and apply it on the set of engineered features for emotion $e$ resulted from previous step. I tuned the hyper-parameters of the classifier is chosen as part of the brute-force search to find the best performance of the multi-label classification task.

## Step 6b: Assignment

As in Step 6a, emotion spans (and their associated emotion labels) were associated with an animate if they occured inside the same semantic role structure.

For every emotion $e$, I use 10-fold cross-validation in Step 4b to evaluate the choice of feature set and classifier in Step 5b, tuning the hyper-parameters of the selected classifier to achieve the best performance. I did a greedy beam search over combinations of feature sets and classifiers, stopping the search when performance improvements fell below 0.5 points of $F_1$.

To calculate the overall $F_1$ score among the ten folds, I use the following formula:

$$\text{overall } F_1 = \frac{\overline{TP}}{\overline{TP} + (\overline{FP} + \overline{FN})/2}$$

where $\overline{TP}, \overline{FP}$, and $\overline{FN}$ represent the average true-positives, false-positives and false-negatives over all the ten folds respectively. With the goal of maximizing $F_1$ score, I repeat steps 5.b and 6.b after modifying/fine-tuning the hyper-parameters of feature engineering tools and classifiers mentioned in the previous steps.

| | All E. | Sig.W. | Setup 1 | | | | | | Setup 2 | | | | | | Setup 3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base | Base | Unsupervised | | | Supervised | | | Unsupervised | | | Supervised | | | Unsupervised | | | Supervised | | |
| Emotion | $F_1$ | $F_1$ | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ |
| anger | 0.30 | 0.69 | 0.84 | 0.92 | 0.88 | 0.73 | 0.87 | 0.80 | 0.67 | 0.74 | 0.70 | 0.61 | 0.94 | 0.74 | 0.19 | 0.47 | 0.27 | 0.26 | 0.64 | 0.37 |
| anticipation | 0.44 | 0.69 | 0.93 | 0.93 | 0.93 | 0.78 | 0.89 | 0.83 | 0.75 | 0.77 | 0.76 | 0.64 | 0.88 | 0.74 | 0.16 | 0.56 | 0.25 | 0.61 | 0.53 | 0.57 |
| disgust | 0.27 | 0.67 | 0.87 | 0.94 | 0.91 | 0.76 | 0.86 | 0.81 | 0.68 | 0.76 | 0.72 | 0.56 | 0.94 | 0.71 | 0.18 | 0.49 | 0.26 | 0.65 | 0.68 | 0.66 |
| fear | 0.38 | 0.73 | 0.87 | 0.98 | 0.92 | 0.77 | 0.93 | 0.85 | 0.73 | 0.81 | 0.76 | 0.68 | 0.98 | 0.80 | 0.23 | 0.51 | 0.31 | 0.73 | 0.42 | 0.53 |
| joy | 0.48 | 0.75 | 0.94 | 0.91 | 0.92 | 0.86 | 0.85 | 0.86 | 0.82 | 0.78 | 0.80 | 0.75 | 0.89 | 0.81 | 0.24 | 0.57 | 0.34 | 0.85 | 0.64 | 0.73 |
| sadness | 0.46 | 0.75 | 0.91 | 0.95 | 0.93 | 0.83 | 0.89 | 0.86 | 0.77 | 0.79 | 0.78 | 0.69 | 0.93 | 0.79 | 0.23 | 0.53 | 0.32 | 0.76 | 0.63 | 0.69 |
| suprise | 0.31 | 0.69 | 0.89 | 0.96 | 0.93 | 0.79 | 0.90 | 0.84 | 0.71 | 0.81 | 0.76 | 0.65 | 0.90 | 0.75 | 0.19 | 0.65 | 0.29 | 0.71 | 0.38 | 0.50 |
| trust | 0.34 | 0.68 | 0.87 | 0.90 | 0.89 | 0.80 | 0.84 | 0.82 | 0.75 | 0.76 | 0.75 | 0.59 | 0.90 | 0.71 | 0.16 | 0.53 | 0.25 | 0.66 | 0.72 | 0.69 |
| Micro Avg. | 0.38 | 0.71 | 0.90 | 0.93 | 0.92 | 0.79 | 0.88 | 83.46 | 0.74 | 0.78 | 0.76 | 0.65 | 0.92 | 0.76 | 0.20 | 0.54 | 0.29 | 0.60 | 0.58 | 0.59 |

Table 6.2: Results of the proposed supervised and unsupervised learning models.

## 6.5 Performance Evaluation on CEN

I applied the described pipeline (6.1) to the 30 annotated chapters of the ABBE corpus which contain a total of 18,725 animate beings. Among all of animate beings, only 2,227 of them are associated with an emotion. I divided my performance evaluations into three tasks listed by the level of difficulty in the following way:

1. I give the system gold standard animate beings (skipping Step 3) and gold standard emotion spans present in the annotated data, and ask the system to produce multi-emotion labels (excluding neutral) for the gold standard spans. This eliminates all possible false positives for emotion span detection, as well as false negatives for the neutral category. This setup tests the effectiveness of the emotion detection component in isolation from the rest of the system.
2. I give the system gold standard animate beings (skipping Step 3), and ask the system to identify both emotion spans and their labels. This tests the performance of the novel components described in this chapter, leaving aside the performance of the animate being detection.
3. I run the complete pipeline without any gold standard inputs, including animate being detection, emotion span detection, emotion span labeling, and span-to-being assignment. This test end-to-end performance.

Instead of breaking out the performance of individual steps, which make comparison between the three setups complicated, I use a single performance measure that measures the ability of each setup (and baseline) to assign the right emotion labels to the right animate beings. Summary scores for all setups and baselines are shown in Table 6.2.

### 6.5.1 Baseline Classifiers

I compared the three setups with two baseline classifiers, both of which used gold standard animate beings. The first is the *All-Emotions* baseline, where each animate being is assigned all possible emotion labels. The second baseline is the *Signal-Words* baseline, where the animate being in the sentence is assigned all emotions for which there exists an emotion word with an associated emotion in the same sentence.

- $B_e^{(0)}(\mathcal{E}_e, \mathcal{W}_e, \mathcal{H}, \mathcal{N}, \mathcal{Q}, \mathcal{C}, \mathcal{R}) =$ true for every emotion $e$. This baseline is referred to by *always-positive* baseline. In this case, precision of each emotion $e$ is equal to its relative frequency, and recall of $e$ is one.
- $B_e^{(1)}(\mathcal{E}_e, \mathcal{W}_e, \mathcal{H}, \mathcal{N}, \mathcal{Q}, \mathcal{C}, \mathcal{R}) = \mathcal{W}_e$ for every emotion $e$. As mentioned before, $\mathcal{W}_e$ determines whether there exists at least one emotional word labeled by $e$ in the sentence containing the reference to animate being.

### 6.5.2 Results

I implemented two solutions (unsupervised and supervised) for the first setup. In the unsupervised solution, I used the unsupervised portion of the proposed emotion detection system to assign the multi-emotion labels to animate-beings. In the supervised solution, I used the multi-label emotion classifier as described in Step 5b. These solutions achieved $F_1$ scores of 0.92 and 0.83 for the unsupervised and supervised cases, respectively.

For the second setup, I first used computed SRL arguments to identify emotion span candidates for every given gold standard animate-being. The recall of this step is 0.65. Then I applied the procedure described in §6.3. The success of this step can be measured in several ways. If measure this strictly, and identify as correct only those spans which exactly match the gold standard, the $F_1$ score for this stage is 0.37. Alternatively, I can compute graded match ratios[1] to handle cases where identified spans overlap with but do not exactly match the gold spans: the $F_1$ score of this step is 0.50. Using this latter approach, I combined the unsupervised portion of the proposed emotion detection system to assign the multi-emotion labels to animate-beings, which achieved an $F_1$ score of 0.76. When combined with the supervised emotion labeling approach, it achieves an $F_1$ of 0.76 as well.

For the third setup, I first identify animate-beings given no annotated data using [Jahan et al., 2018a]. Then, I used computed SRL arguments to identify emotion span candidates for every identified animate-being. Considering graded match ratios of identified emotion spans with gold standard ones, the recall of this step is 0.55. If I apply strict match, the recall is 0.46. Finally, I used the unsupervised portion of the proposed emotion detection system to assign the multi-emotion labels to animate-beings. The $F_1$ score is this case was poor, at 0.29. However, when applying the supervised model, I achieve a much better $F_1$ of 0.59.

## 6.6 Unsolved Challenges and Future Work

The work described presents several opportunities for improvement, which I describe below.

[1]Given two strings $s$ and $t$ if $S$ and $T$ denote the set of tokens in $s$ and $t$ respectively, I define graded match ratio is $|S \cap T|/|S \cup T|$.

First, built into the approach is the assumption that animate beings and their associated emotions are related by being semantic arguments to the same verb. However 658 of the animate beings were not inside an argument to a verb as computed by the SRL. Table 6.3 shows a few examples of the missed SRL tags.

Table 6.3: Examples of inaccuracy of semantic role labeling.

| Inaccurate Semantic Role Labeling Example | Missed Animate being |
|---|---|
| He [V: seemed] [ARG1: not at all displeased] . | He |
| Well , ' he cried , ' and [ARG2: whose fault] [V: was] [ARG1: it] but mine ? | mine |
| Beatrice must grow used to the idea of marriage and [ARGM-MOD: must] [V: be] gradually accustomed to the daily companionship of San Miniato . | San Miniato |
| [ARGM-DIS: Of course] [ARG0: I] [V: think] [ARG1: so] , " quoth John stoutly . | John |

The rest of animate being referring expressions (18,067) were contained within 24,320 separate SRL arguments; this means that each animate being corresponded to on average 1.35 SRL arguments. This is because, in complex and long sentences, SRL identifies long arguments in a way that in many cases, each argument contains multiple other arguments (nested arguments). Therefore, SRL cannot filter out the arguments (emotion spans) that are neither directly nor indirectly related to an animate-being in a long sentence. As a result, SRL gives too many emotion span candidates for a single animate-being which increases false positives and lowers precision substantially. This one-to-many relationship complicates the identification of the related emotion spans. Further, my SRL implementation did not recognize some verbs; this reduced performance even more. A more sophisticated way of identifying the correspondence between identified emotion spans and animate beings (perhaps not relying on SRL) would perhaps improve performance.

Second, Among the 18,725 animate beings, only 2,227 of them are associated with an emotion. This imbalance between the positive and the negative classes makes learning the modeling challenging; a more balanced dataset, or more positive examples, would likely increase performance, at least of the supervised approach.

Third, sometimes the NRC lists multiple secondary labels besides the primary label to a word that are not all necessarily conveyed in a specific text. Also, my future work includes expanding the coverage of the revised NRC lexicon for 8-class Plutchik's emotions.

Fourth, the tf-idf vector space model cannot be used to cancel noise very effectively (in contrast to NMF), because of the smallness of the data: with only thirty chapters, the popularity of various words are close to each other, and does not present a clear cutpoint in the distribution beyond which to discard words.

Finally, one possible way to further improve the performance is to increase the amount of data that requires more annotated data based on different psychological models and the related emotion lexicons.

## 6.7 Contributions

I designed a system that detects emotion of animate beings using Revise -NRC lexicon for Plutchik's psychological model (eight emotions), and trained and tested on the ABBE of thirty chapters of the Corpus of English Novels. When isolated from the effects of the animate being detection system, the emotion detection and association portions of the work achieved an overall micro $F_1$ of 0.76.

Chapter 7

# Summary and Contributions

In my dissertation, I focused on major efforts that are critical to the field of narrative and language understanding.

In Chapter §2, I reviewed literature related to psychological theories in great detail, emotion lexicons, emotion datasets, emotion detection approaches, and language resources for animate beings. Based on these reviewed I discovered that prior work on emotion detection suffered from noisy, incomplete, and unreliable data, as well as unreproducible results [Zad et al., 2021d].

Then, I proceeded to identify the most promising piece of prior work and reimplement it in a way that fully evaluated and reproducible. This resulted in my first paper , in 2020 in the ACL-hosted NUSE workshop, which demonstrated a state-of-the-art system for detection emotions in narrative text, and which is a model of reproducible science [Zad and Finlayson, 2020]. In Chapter §3, I identified a high performing approach to emotion recognition in narrative text [Kim et al., 2010] and carefully reimplemented and characterized the technique, exploring a design space of three different noise cancellation or dimension reduction techniques (NMF, PCA, or LDA), exploring various hyperparameter settings. My experiments indicated that NMF performed best, with an overall $F_1$ of 0.809. In the course of my investigation I clarified numerous implementational issues of the work reported by Kim et al. [2010], as well as made some improvements to WordNet Affect (WNA) by adding new terms manually and using Wordnet similarity relations. This work suggests several promising future directions for improving the work, including careful annotation of a larger corpus, and augmenting WNA or similar lexicons to provide improved coverage of emotion terms. I released my code and data[1].

[1]Code and data can be downloaded from `https://doi.org/10.34703/gzx1-9v95/03RERQ`

I then turned to a related problem that had revealed itself during this work, namely, that the emotion lexicon I was using, the NRC EmoLex Lexicon, contained a number of seemingly incorrect or biased entries. This lexicon is one of the most widely used language resources in emotion detection, but on close inspection a number of problems emerge. For example, words that should in most contexts be emotionally neutral, with no affect (e.g., stone labeled as ANGER, mountain as ANTICIPATION), were associated with emotional labels that are inaccurate, nonsensical or, at best, highly contingent and context-dependent, while other entries were downright pejorative (e.g., lesbian labeled as DISGUST and SADNESS). I along with my undergraduate research assistant Joshua Jimenez, developed a semi-automatic procedure for correcting the NRC (resulting in an updated resource called the Revised NRC), which not only greatly reduced the number of problematic entries but actually increased the performance of a system that used the NRC. Bias in AI and NLP systems has become a major issue of concern in recent years, and this work makes a clear contribution to that area. This paper was published in the 5th Workshop on Online Abuse and Harms in 2021 [Zad et al., 2021c]. In Chapter §4, I noted three categories of error in the popular NRC emotion lexicon, including a large number of seemingly biased entries. I developed and applied a semi-automatic procedure to generate three different corrected version of the NRC, and showed via experiment that these new versions improved the performance of an existing emotion detection system. This chapter shows the utility of careful error checking of lexical resources, especially with attention to correcting for unintended biases. Finally, I release the revised resource and my code to enable other researchers to reproduce and build upon results[2].

My fourth paper, currently under review at LREC, concerns the construction of the dataset that captures the phenomena in question. In this work, Josh and I laboriously double-annotated 30 narrative texts comprising 134k words for experienced emotion and

[2] https://doi.org/10.34703/gzx1-9v95/PO3YGX

the associated animate beings. We call this corpus the ABBE corpus: Animate Beings Being Emotional. This work shows all the hallmarks of a carefully designed and executed linguistic annotation study, where the phenomenon under study is carefully and exactly defined, the annotation scheme is clearly explained, the data is double-annotated with high inter-annotator agreement, and adjudicated for correctness. This corpus clearly fills a gap in the field of emotion detection, and I expect it will be a resource used by many in the field for years to come [Zad et al., 2022]. My contributions in Chapter §5 are threefold. First, I note the importance of the experiencer to the idea of emotion, and point out that all prior corpora of annotation emotion omit this information. Second, I define an annotation scheme that captures the experiencer of an emotion (an animate being), based on prior work on emotion classification and animate being detection. Finally, I provide the ABBE corpus, a collection of 30 texts (134.5k tokens) annotated for 2,010 emotion spans associated with 2,227 animate beings, annotated with excellent inter-annotator agreement. I release the ABBE data for other researchers to use in their work[3].

My fifth paper uses the ABBE corpus to demonstrate the first system that both identifies experienced emotions as well as the animate beings that are experiencing those emotions. Building on my prior work on emotion detection, as well as the Revised NRC, I developed an NLP system for carrying out the four steps necessary to solve this task: (1) identifying emotional spans, (2) labeling those emotion spans, (3) identifying animate beings, and (4) associating animate beings with the experienced emotions In Chapter §6. Step #3 was work done by Labiba Jahan at Cognac Lab, but Steps 1, 2, and 4 were all novel to my system. I explored many possible system architectures, demonstrating great perseverance in exploring different configurations and approaches to solving the problem. One thing that was surprising about my results was how difficult the task turned out to be: when the system is isolated from animate being detection, overall performance is 0.76

[3]Data and code may be downloaded from `https://anonymized.url`.

$F_1$. When detection of animate beings is included, performance drops to 0.6 $F_1$. It points to the significant challenges in this task which result from the subtlety and subjectivity of identifying emotion, and the difficulty of connecting an emotion to its experiencer, especially in complex literary narrative texts.

# BIBLIOGRAPHY

Isidoros Perikos and Ioannis Hatzilygeroudis. Recognizing emotion presence in natural language sentences. In *International conference on engineering applications of neural networks*, pages 30–39, Berlin, Heidelberg, 2013. Springer. URL https://doi.org/10.1007/978-3-642-41016-1_4.

Matthew W Kreuter, Melanie C Green, Joseph N Cappella, Michael D Slater, Meg E Wise, Doug Storey, Eddie M Clark, Daniel J O'Keefe, Deborah O Erwin, Kathleen Holmes, et al. Narrative communication in cancer prevention and control: a framework to guide research and application. *Annals of Behavioral Medicine*, 33(3): 221–235, 2007.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791, 2012. URL https://doi.org/10.1136/amiajnl-2011-000784.

Haoruo Peng, Daniel Khashabi, and Dan Roth. Solving hard coreference problems. *arXiv preprint arXiv:1907.05524*, 2019. URL https://arxiv.org/abs/1907.05524.

Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4): 371–388, 2005. URL https://doi.org/10.1016/j.neunet.2005.03.002.

Samira Zad, Maryam Heidari, James H Jones, and Ozlem Uzuner. A survey on concept-level sentiment analysis techniques of textual data. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0285–0291. IEEE, 2021a.

Samira Zad, Maryam Heidari, H James Jr, and Ozlem Uzuner. Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0255–0261. IEEE, 2021b.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014): System Demonstrations*, pages 55–60, Baltimore, MD, U.S., 2014.

Ebba Cecilia Ovesdotter Alm. *Affect in *Text and Speech*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 2008.

Carlo Strapparava and Alessandro Valitutti. WordNet Affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1083–1086, Lisbon, Portugal, 2004.

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA, 1998a.

Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:433–459, 2010. URL `https://doi.org/10.1002/wics.101`.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. URL `https://doi.org/10.1038/44565`.

Labiba Jahan, Geeticka Chauhan, and Mark A Finlayson. A new approach to animacy detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2018a.

Labiba Jahan, W Victor H Yarlott, Rahul Mittal, and Mark A Finlayson. Confirming the generalizability of a chain-based animacy detector. In *AI4Narratives@ IJCAI*, pages 43–46, 2020a.

Joshua Eisenberg and Mark Finlayson. A simpler and more generalizable story detector using verb and character features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2708–2715, 2017.

Mark Alan Finlayson. Collecting semantics in the wild: The story workbench. In *Proceedings of the AAAI Fall Symposium on Naturally Inspired Artificial Intelligence*, pages 46–53. Arlington, VA, 2008.

Mark A Finlayson. The Story Workbench: An extensible semi-automatic text annotation tool. In *Proceedings of the 4th Workshop on Intelligent Narrative Technologies (INT4)*, pages 21–24. Stanford, CA, 2011.

Robert Plutchik. A general psychoevolutionary theory of emotion. In Robert Plutchik, editor, *Theories of Emotion*, pages 3–33. Elsevier, Amsterdam, Netherlands, 1980. URL https://doi.org/10.1016/B978-0-12-558701-3.50007-7.

Robert Plutchik. Emotions and imagery. *Journal of Mental Imagery*, 8:105–111, 1984.

Robert Plutchik. *The psychology and biology of emotion.* HarperCollins College Publishers, 1994.

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

Samira Zad, Joshuan Jimenez, and Mark A. Finlayson. The abbe corpus: Animate beings being emotional. 2022.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013. URL https://arxiv.org/abs/1308.6242.

Saif M Mohammad and Peter D Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013. URL https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, 2010.

J. Bernard, editor. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia, 1986.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331 (6014):176–182, 2011.

Feride Savaroğlu Tabak and Vesile Evrim. Comparison of emotion lexicons. In *Proceedings of the 13th Interntaional Symposium on Smart Microgrids for Sustainable Energy Sources Enabled by Photonics and IoT Sensors (HONET-ICT)*, pages 154–158, Nicosia, Cyprus, 2016. doi: 10.1109/HONET.2016.7753440.

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855, 2017.

S Lovelyn Rose, R Venkatesan, Girish Pasupathy, and P Swaradh. A lexicon-based term weighting scheme for emotion identification of tweets. *International Journal of Data Analysis Techniques and Strategies*, 10(4):369–380, 2018.

Young-Jun Lee, Chan-Yong Park, and Ho-Jin Choi. Word-level emotion embedding based on semi-supervised learning for emotional classification in dialogue. In *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4, Kyoto, Japan, 2019. doi: 10.1109/BIGCOMP.2019.8679196.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. The lilah emotion lexicon of croatian, dutch and slovene. In *Proceedings of the 3rd Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 153–157, Barcelona, Spain (Online), 2020.

Samira Zad, Maryam Heidari, James H Jr Jones, and Ozlem Uzuner. Emotion detection of textual data: An interdisciplinary survey. In *Proceedings of the IEEE World AI IoT Congress (AIIoT 2021)*, Seattle, WA, 2021d.

Amy M. Schmitter. 17th and 18th Century Theories of Emotions. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

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

William James. Discussion: The physical basis of emotion. *Psychological review*, 1(5): 516, 1894.

Gary R VandenBos. *APA dictionary of psychology.* American Psychological Association, 2007.

Michael W Eysenck, Andrew W Ellis, Earl B Hunt, and Philip Nicholas Ed Johnson-Laird. *The Blackwell dictionary of cognitive psychology.* Basil Blackwell, 1994.

Carroll E Izard. Basic emotions, relations among emotions, and emotion-cognition relations. *US: American Psychological Association*, 1992.

Klaus R Scherer. Neuroscience projections to current debates in emotion psychology. *Cognition & Emotion*, 7(1):1–41, 1993.

Rafael A Calvo and Sunghwan Mac Kim. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543, 2013. URL https://doi.org/10.1111/j.1467-8640.2012.00456.x.

Keith Oatley and Philip N Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1(1):29–50, 1987. URL https://doi.org/10.1080/02699938708408362.

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992a. URL https://doi.org/10.1080/02699939208411068.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086, 1987. URL https://doi.org/10.1037/0022-3514.52.6.1061.

W Gerrod Parrott. *Emotions in Social Psychology: Essential Readings*. Psychology Press, London, 2001.

Jaak Panksepp, Brian Knutson, and Douglas L Pruitt. Toward a neuroscience of emotion. In *What develops in emotional development?*, pages 53–84. Springer, Boston, MA, 1998. URL https://doi.org/10.1007/978-1-4899-1939-7_3.

Carroll E Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, 2(3):260–280, 2007.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. URL https://doi.org/10.1037/h0077714.

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

Cynthia M Whissell. The dictionary of affect in language. In *The measurement of emotions*, pages 113–131. Elsevier, 1989.

Hugo Lövheim. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78(2):341–348, 2012. URL https://doi.org/10.1016/j.mehy.2011.11.016.

Andrew Ortony, Gerald L Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK, 1990.

Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent Müller, editors, *Cognitive Behavioural Systems*, pages 144–157. Springer, Berlin, 2012. Published as Volume 7403, Lecture Notes in Computer Science (LNCS).

Robert Plutchik. 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, 2001a.

Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–30, 2018. URL `https://doi.org/10.1145/3161414`.

Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136, Berlin, Germany, 2013. URL `https://doi.org/10.1007/978-3-642-37256-8_11`.

Paul Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992b. URL `https://doi.org/10.1037/0033-295X.99.3.550`.

Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. URL https://doi.org/10.1037/0003-066X.48.4.384.

George Langroudi, Anna Jourdanous, and Ling Li. Music emotion capture: Sonifying emotions in eeg data. In *Symposium on Emotion Modeling and Detection in Social Media and Online Interaction*, pages 1–4, Liverpool, UK, 2018.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70, Los Angeles, CA, 2010.

James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999. URL https://doi.org/10.1037/0022-3514.76.5.805.

Robert Ed Plutchik and Hope R Conte. *Circumplex Models of Personality and Emotions.* American Psychological Association, Washington, DC, 1997.

Gerardo Maupome and Olga Isyutina. Dental students' and faculty members' concepts and emotions associated with a caries risk assessment program. *Journal of Dental Education*, 77(11):1477–1487, 2013.

Gerald L Clore and Andrew Ortony. Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335–343, 2013.

Bas R Steunebrink, Mehdi Dastani, and John-Jules Ch Meyer. The occ model revisited. In *Proc. of the 4th Workshop on Emotion and Computing*. Association for the Advancement of Artificial Intelligence, 2009.

Isidoros Perikos and Ioannis Hatzilygeroudis. Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, 51:191–201, 2016.

Yosephine Susanto, Andrew G Livingstone, Bee Chin Ng, and Erik Cambria. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102, 2020.

DA Medler, A Arnoldussen, JR Binder, and MS Seidenberg. The Wisconsin Perceptual Attribute Ratings (WPAR) database, 2005. Retrieved from `http://www.neuro.mcw.edu/ratings` on April 23, 2020.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic Inquiry and Word Count (LIWC) Software, 2001.

Anil Bandhakavi, Nirmalie Wiratunga, P Deepak, and Stewart Massie. Generating a word-emotion lexicon from #emotional tweets. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 12–21, Dublin, Ireland, 2014.

Jacopo Staiano and Marco Guerini. DepecheMood: A lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

Ema Kušen, Giuseppe Cascavilla, Kathrin Figl, Mauro Conti, and Mark Strembeck. Identifying emotions in social media: Comparison of word-emotion lexicons. In *Proceeedings of the 5th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 132–137, Prague, Czech Republic, 2017. doi: 10.1109/FiCloudW.2017.75.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. DepecheMood++: A bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. doi: 10.1109/TAFFC.2019.2934444.

Margaret M Bradley and Peter J Lang. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, Gainseville, FL, 1999.

Philip James Stone, Dexter Colboyd Dunphy, Daniel M Ogilvie, and Marshall S Smith. *The General Inquirer: a Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966.

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

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

Christiane Fellbaum. Towards a representation of idioms in wordnet. In *Proceedings of the Workshop on the Use of WordNet in Natural Language Processing Systems*, pages 52–57, Montreal, Canada, 1998b.

James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic inquiry and word count: LIWC [computer software], 2007. URL `https://liwc.net`.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26, 2018.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

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

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

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*, 2019.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1073.

Cecilia Ovesdotter Alm. Characteristics of high agreement affect annotation in text. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 118–122, Uppsala, Sweden, 2010.

Samira Zad and Mark Finlayson. Systematic evaluation of a framework for unsupervised emotion recognition for narrative text. In *Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 26–37, Online, 2020. doi: 10.18653/v1/2020.nuse-1.4. URL https://www.aclweb.org/anthology/2020.nuse-1.4.

Samira Zad, Maryam Heidari, Parisa Hajibabaee, and Masoud Malekzadeh. A survey of deep learning methods on semantic similarity and sentence modeling. In *2021*

*IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0466–0472. IEEE, 2021e.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://aclanthology.org/S17-2088.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas Pykl, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, 2020.

Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2092.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *CICLing*, 2015.

Timothy Tackett. *The Coming of the Terror in the French Revolution*. Harvard University Press, Cambridge, MA, U.S., 2015.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. EmoTxt: a toolkit for emotion recognition from text. In *Proceedings of the Seventh International Conference on Affective Computing and Intelligent: Interaction Workshops and Demos (ACIIW 2017)*, pages 79–80, San Antonio, TX, 2017. URL https://doi.org/10.1109/ACIIW.2017.8272591.

Zhi Teng, Fuji Ren, and Shingo Kuroiwa. Emotion recognition from text based on the rough set theory and the support vector machines. In *Proceedings of the 2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 36–41, Beijing, China, 2007.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. Emotion recognition from text based on automatically generated rules. In *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop*, pages 383–392, Shenzhen, China, 2014.

Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, and Andrés Montoyo. Ua-zbsa: a headline emotion classification through web information. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 334–337, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S07-1072.

Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, Fortaleza, Ceara, Brazil, 2008. URL https://doi.org/10.1145/1363686.1364052.

Saima Aman and Stan Szpakowicz. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 312–318, Hyderabad, India, 2008.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics: Volume 1*, pages 881–888, Manchester, UK, 2008.

Colin Cherry, Saif M Mohammad, and Berry de Bruijn. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5:BII–S8933, 2012. URL `https://doi.org/10.4137/BII.S8933`.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592, Washington, DC, U.S., 2012. IEEE. URL `https://doi.org/10.1109/SocialCom-PASSAT.2012.119`.

Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. *arXiv preprint arXiv:1804.00831*, 2018.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, 93:133–142, 2017. URL https://doi.org/10.1016/j.patrec.2016.12.009.

Parisa Hajibabaee, Masoud Malekzadeh, Maryam Heidari, Samira Zad, Ozlem Uzuner, and James H Jones. An empirical study of the graphsage and word2vec algorithms for graph multiclass classification. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0515–0522. IEEE, 2021.

Masoud Malekzadeh, Parisa Hajibabaee, Maryam Heidari, Samira Zad, Ozlem Uzuner, and James H Jones. Review of graph neural network in text classification. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0084–0091. IEEE, 2021.

Maryam Heidari, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, and James H Jones. Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0103–0109. IEEE, 2021a.

Maryam Heidari, Samira Zad, and Setareh Rafatirad. Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–6. IEEE, 2021b.

Maryam Heidari, Samira Zad, Brett Berlin, and Setareh Rafatirad. Ontology creation model based on attention mechanism for a specific business domain. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–5. IEEE, 2021c.

Andrea Chiorrini, Claudia Diamantini, Alex Mircoli, and Domenico Potena. Emotion and sentiment analysis of tweets using bert. 2021.

Md Shad Akhtar, Asif Ekbal, and Erik Cambria. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1):64–75, 2020.

Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. In *Informatics*, volume 8, page 19. Multidisciplinary Digital Publishing Institute, 2021.

Xiaojun Quan, Qifan Wang, Ying Zhang, Luo Si, and Liu Wenyin. Latent discriminative models for social emotion detection with emotional dependency. *ACM Transactions on Information Systems (TOIS)*, 34(1):1–19, 2015.

Valentina Sintsova, Claudiu-Cristian Musat, and Pearl Pu. Fine-grained emotion recognition in olympic tweets based on human computation. In *4th Workshop on computational approaches to subjectivity, sentiment and social media analysis*, number CONF, Atlanta, Georgia, 2013.

Samuel Bowman and Harshit Chopra. Automatic animacy classification. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, pages 7–10, 2012a.

Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2015.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics*, 39(4):885–916, 2013.

Labiba Jahan, Rahul Mittal, W. Victor Yarlott, and Mark A. Finlayson. A straightforward approach to narratologically grounded character identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6089–6100, Barcelona, Spain (Online), December 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.536. URL `https://www.aclweb.org/anthology/2020.coling-main.536`.

Hendrik De Smet. Corpus of English novels, 2008a. `https://perswww.kuleuven.be/~u0044428/`.

Samuel Bowman and Harshit Chopra. Automatic animacy classification. In *North American Association for Computational Linguistics - Human Language Technologies*

*(NAACL-HLT) Student Research Workshop*, 2012b.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Ian Beaver. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44:387–419, 2010.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. Animacy encoding in English: Why and how. In *Proceedings of the Workshop on Discourse Annotation*, pages 118–125, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-0216.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696, Montreal, Canada, 2009. URL https://doi.org/10.1145/1553374.1553463.

Christos Boutsidis and Efstratios Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008. URL https://doi.org/10.1016/j.patcog.2007.09.010.

Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932, San Jose, CA, 2007.

Taner Danisman and Adil Alpkocak. Feeler: Emotion classification of text using vector space model. In *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, volume 1, page 53, Aberdeen, Scotland, 2008.

Ameeta Agrawal and Aijun An. Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 346–353, Macau, China, 2012. doi: 10.1109/WI-IAT.2012.170.

Samar Al-Saqqa, Heba Abdel-Nabi, and Arafat Awajan. A survey of textual emotion detection. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, pages 136–142, Los Alamitos, CA, 2018. doi: 10.1109/CSIT.2018.8486405.

iWeb. The iWeb Corpus, 2021. `https://www.english-corpora.org/iweb/`. Last accessed on April 25, 2021.

Mark Davies and Jong-Bok Kim. The advantages and challenges of "big data": Insights from the 14 billion word iWeb corpus. *Linguistic Research*, 36(1):1–34, 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*, 2016.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the*

*Association for Computational Linguistics*, 6:587–604, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67, Washington, DC, 2010.

Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, Beijing, China, 2011.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Jimena Olveres, Mark Billinghurst, Jesus Savage, and Alistair Holden. Intelligent, ex-

pressive avatars. In *Proceedings of the First Workshop on Embodied Conversational Characters*, pages 47–55, 1998.

Erik T Mueller. Natural language processing with thought treasure, 1998.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48, 2019.

Charles Darwin and Phillip Prodger. *The Expression of the Emotions in Man and Animals*. Oxford University Press, Oxford, UK, 1998.

William James. *The Principles of Psychology*. Henry Holt and Company, New York, 1890.

Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

Robert Plutchik. Integration, differentiation, and derivatives of emotion. *Impressum Evolution and Cognition*, 7(2):114–126, 2001b.

Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. A new approach to animacy detection. In *the 27th International Conference on Computational Linguistics (COLING)*, pages 1–12, Santa Fe, NM, 2018b. URL `http://aclweb.org/anthology/C18-1001`.

David Herman. *Cognitive narratology*. de Gruyter, 2014.

Seymour Chatman. Story and discourse. narrative structure in fiction and film, cornell, 1986.

Uri Margolin. Individuals in narrative worlds: An ontological perspective. *Poetics Today*, 11(4):843–871, 1990.

Haji Binali, Chen Wu, and Vidyasagar Potdar. Computational approaches for emotion detection in text. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, pages 172–177. IEEE, 2010.

Nilesh M Shelke. Approaches of emotion detection from text. *International Journal of Computer Science and Information Technology*, 2(2):123–128, 2014.

Lea Canales and Patricio Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43, 2014.

Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. Emotion detection: a technology review. In *Proceedings of the XVIII international conference on human computer interaction*, pages 1–8, 2017.

Hendrik De Smet. Corpus of english novels, 2008b. URL `https://perswww.kuleuven.be/~u0044428/`.

Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255, 2019.

Ian Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.

# VITA

## SAMIRA ZAD

| | |
|---|---|
| 2017–2022 | Ph.D., Computer Science<br>Florida International University<br>Miami, Florida, USA |
| 2021-2022 | Three best paper awards in peer-reviewed IEEE conferences |
| 2019–2021 | Mentor<br>Joshuan Jimenez,<br>Undergrad. and Graduate Research Assistant<br>Co-led the Artificial Intelligence team<br>SparkDev<br>Florida International University<br>Miami, Florida, USA |
| 2018–2022 | Graduate Research Assistant<br>Cognition, Narrative and Cultural Laboratory<br>Florida International University<br>Miami, Florida, USA |
| 2021 | Scientific Journal and Conference Paper Reviewer<br>International Journal of Data Science and Analytics<br>Online |
| 2018 | Full scholarship<br>Center for advancing Education and studies on<br>Critical Infrastructure Resilience (CAESCIR)<br>for my PhD years 2 to 5 |
| 2018 | Grace Hopper Scholarship<br>Florida International University |
| Summer, 2018 | Summer Research Intern<br>U.S. Department of Homeland Security<br>University of Illinois at Urbana-champaign<br>Urbana-champaign, Illinois, USA |
| 2017 | One year Graduate Assistantship<br>School of Computing and Info. Sciences<br>Florida International University |

| | |
|---|---|
| 2014–2016 | M.S., Computer Science<br>Northeastern Illinois University<br>Chicago, Illinois, USA |
| 2000–2005 | B.Sc., Mathematics<br>University of Kashan<br>Kashan, Iran |

PUBLICATIONS

Samira Zad and Mark A. Finlayson (under review) Detecting the Emotions of Animate Beings in Narrative. Submitted to the ACL Rolling Review Process for publication in an ACL main conference in 2022.

Samira Zad, Joshuan Jimenez, Mark A. Finlayson (under review) The ABBE Corpus: Animate Beings Being Emotional. Submitted to the 13th Language Resources and Evaluation Conference (LREC), to be held in Marseille, France, 2022.

Samira Zad and Mark A. Finlayson (2020). Systematic Evaluation of a Framework for Unsupervised Emotion Recognition for Narrative Text. In Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE) co-located with ACL, Seattle, USA (Online). (pp. 26-37)

Samira Zad, Joshuan Jimenez, Mark A. Finlayson (2021). Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon. In Proceedings of the 5th Workshop on Online Abuse and Harms co-located with ACL, Bangkok, Thailand (Online). (pp. 102-113).

Samira Zad, Maryam Heidari, James H Jr Jones, Ozlem Uzuner (2021). Emotion Detection of Textual Data: An Interdisciplinary Survey. In Proceedings of the International Conference 2021 IEEE World AI IoT Congress (AIIoT), Seattle, USA (Online), (pp. 225-261).

Samira Zad, Maryam Heidari, James H Jr Jones, Ozlem Uzuner (2021). A Survey on Concept-Level Sentiment Analysis Techniques of Textual Data. In Proceedings of the International Conference 2021 IEEE World AI IoT Congress (AIIoT), Seattle, USA (Online), (pp. 285-291).