AFFECTIVE ANALYSIS OF TEXT IN TWEETS

BY

ABHISHEK NARWEKAR

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Associate Professor Roxana Girju

Abstract

Affective computing is the study and development of devices that can recognize emotions through various modes such as video, audio and text automatically. In this thesis, I focus on the problem of affective computing in short texts, in particular, tweets. With the evolution of social media in the recent years, there has been a rapid growth of interactions that take occur online, which are expressive in terms of emotion. Internet users today have several diverse methods of being expressive through text, such as by using abbreviations, emoticons and hashtags. I use traditional lexical features and word embeddings to extract semantic and lexical information from the input text. I develop models ranging from linear and tree-based models to deep neural networks to perform emotion detection on Tweets. I create an ensemble of these methods to make my final predictions. I evaluate the ensemble on the SemEval 2018 dataset containing intensity and class annotations for emotions in tweets. I finally perform an error analysis of these algorithms and highlight potential areas of improvement.

Table of Contents

Chapter 1 Introduction

When dealing with people, remember you are not dealing with creatures of logic, but with creatures of emotion.

– Dale Carnegie

Emotions are fascinating. As humans, we express express emotions through affects - facial, vocal or gestural manifestations of our cognitive state. The study and development of devices that can recognize these emotions constitutes the field of **affective computing**. Written text contains a wealth of emotional expression. A novelist's mastery over their craft can often be seen by the way they captivate their readers using the emotions in their words. With the evolution of social media in the recent years, there has been a rapid growth of interactions that take occur online, which are expressive in terms of emotion. Internet users today have several diverse methods of being expressive through text, such as by using abbreviations, emoticons and hashtags[1].

In this thesis, I consider the task of emotion detection in short expressive texts, in particular: tweets. My major contributions are as follows:

- I perform a thorough literature review spanning the fields of psychology and computational linguistics to explore the origins and theories of emotion and computational methods of modeling emotion.
- I develop traditional and distributed models on traditional lexical features and word embeddings to perform emotion detection on Tweets. I create an ensemble of these models to perform my final predictions.
- I evaluate the ensemble on the emotion-annotated SemEval 2018 dataset comprising of about 8000 annotated points across four emotions.
- I perform an error analysis over various algorithms and identify potential sources of improvement to the model.

Let us begin by looking at some important applications of automated emotion detection in text.

1.1 APPLICATIONS OF EMOTION DETECTION

Automated affect recognition systems have several diverse useful applications in academic as well as industrial settings. Here, I briefly discuss some major use cases.

1. Business and Commercial Applications

    - Tracking sentiment towards people and organizations can help understand the public opinion about influential individuals like politicians, actors, sportspersons [75]. Tracking emotions in reviews can be used in applications such as recommender sytems, etc.
    - By incorporating emotion in text, it would be possible to develop more natural text-to-speech systems [36].
    - Coming to the task of generating emotion-rich text, it may be possible to develop systems that assist users express their emotions more effectively [57]. For instance, the system may recommend more expressive phrases to an author, or more articulate ways to convey the emotions in an email, etc.

2. Better Human Computer Interaction

    - One may construct dialogue systems that adapts its behaviour based on the emotional state of the user [102]. As an example, consider the benefits of performing emotion detection in automated customer relations systems during interaction of customers with the system. The system could, for instance, detect anger and redirect the customer to a human if required [16], thereby enhancing user experience.
    - Studies suggest that learning is accelerated when the students are in a positive state of mind [56]. Based on this hypothesis, I may construct a tutoring system that can manage the emotional state of the student.
    - In a similar vein, conversational agents such as Woebot [35] may perform duties similar to a psychiatrist by recognizing the emotional state of a user and improving it if it isn't well.

3. General Text Analysis

    - Researchers can analyze the flow of emotions in a piece of text, such as a novel or a news article [15]. If I construct a model that associates context with a character in a novel, it would be a very interesting idea to study the emotional states of various characters in a story and their relationship with other characters in terms of the emotions they feel because of them
    - It would be very interesting to study aspects of how humans use language itself: it would be possible to study contexts in which humans express themselves and use emotion-rich language [53] - for instance, to coerce or persuade others, etc.

- An artistic feature of language is the use of figurative language. Authors may use figurative phrases such as "*sent a chill down my spine*" to express fear. Such expressions are effective at painting a vivid picture in the mind of the reader. However, they may be devoid of any emotion-bearing keywords. It would be a very interesting challenge to reliably identify the emotional content in such expressions.

1.2 ORIGINS OF EMOTION

Before I delve into computational methods to infer emotions in text, it is insightful to ask: why do I feel emotions? How do emotions originate? Clearly, they are a consequence of the circumstances around us (termed as *stimuli* in literature on psychology), but what processes lead from a stimulus to the actual manifestation of the emotion?

The earliest work on the purpose of emotions was by Darwin (1872) [27]. In their book, *The Expression of Emotions in Man and Animals*, the authors hypothesize that emotions serve the purpose of improving the reproductive fitness of a species. In other words, emotions assist species to survive and reproduce. Fear stands out as a prime example. If an organism detects danger to itself in the form of physical harm, due to fear, it is motivated to take quick actions that take it out of harm's way. For instance, the presence of fear may result in a fight-or-flight response by the organism.

However, there is significant disagreement in the Psychology community on the subject of whether animals other humans than are capable of feeling emotions (Ekman, 1998 [31]). This problem is especially difficult since there isn't yet a way to detect the presence of emotions in animals (De Waal, 2016 [28]). While animals may demonstrate behaviour similar to humans upon receiving rewards or being threatened, there isn't conclusive evidence on whether those responses are due to the manifestation of *emotions*. Skeptics of the animals-with-emotions theory point out that alternate hypotheses may explain their behaviour. For instance, one explanation for animal behaviour upon reward is simply the neural activity in the reward centers of their brain. Although human bodies are physiologically very similar to the body of other mammals, their brains are orders of magnitude more complex than those of mammals in terms of neural connectivity. Since emotions are the result of complex mental processes involving appraisal of situations with respect to situations and objectives, brains of other animals may not be capable of generating emotions (Barrett, 2017 [12]).

With this in mind, it is interesting to ask: how does the human brain generate emotion? Let us briefly look at some theories that explain this phenomenon.

1.3 THEORIES OF EMOTION

There are several theories explaining the process that occurs after experiencing a stimulus that leads to an emotional response. The most popular theories of emotion can be categorized into two main classes: physiological and cognitive. Physiological theories hypothesize that responses in the body result in emotions. On the other hand, cognitive theories suggest that thoughts resulting from our interaction with the surroundings are critical to forming emotions. Figure 1.1 contains an intuitive and visual description of four major theories of emotion for an example of a situation that results in *fear*. Let us look overview some of these theories breifly.

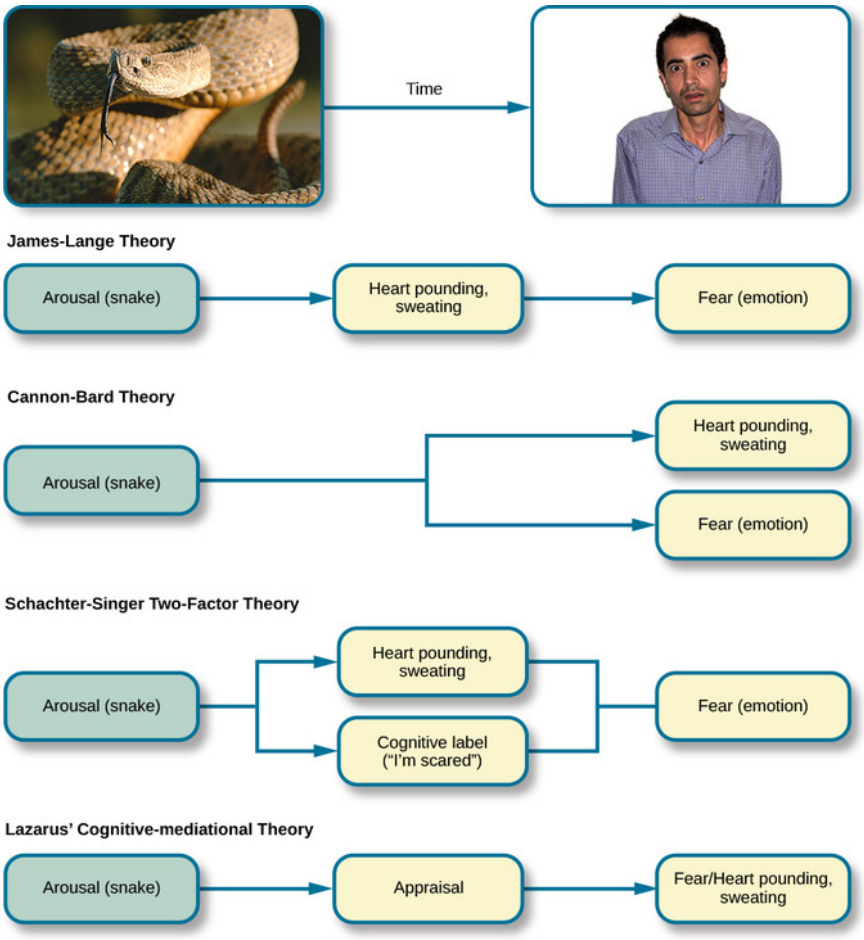


Figure 1.1: A visual comparison of the theories of emotion (source)

1.3.1 Physiological Theories

James-Lange theory (Lange and James, 1922 [54])
Proposed independently by psychologist William James and physiologist Carl Lange, this theory is one of the major psychological theories of emotion. It suggests that emotions are a result of the body's physiological reactions to stimuli.

In other words, this theory proposes that when we come in contact with an external stimulus, it results in a physiological reaction. According to this theory, the emotional response depends on our *interpretation* of those physiological reactions. For instance, if one observes a snake while walking down a dark road, unless they are expert at handling snake, their body will likely produce a physiological response, such as sweat, elevated heartbeat and trembling. According to Lange and James, they will interpret these physical reactions in a way that leads to the conclusion that they are frightened. In other words, one is afraid because they are trembling, and not the other way around.

The James-Lange theory has some rather counterintuitive implications. It says, for instance, that we don't cry because we are sad. Rather, crying is a physiological response to an experience, due to which we feel sad. It has faced some criticism:

- The mapping between physiological responses and emotions is not injective (one-to-one). The same physiological response may co-occur with different emotions. For instance, quickening of the heartbeat due to exercise doesn't necessarily result in fear.
- Physiological responses are sometimes too slow, while emotions can be instantaneous, such as those associated with fear.
- Later work in the area of neuroscience showed that that both animals and humans having experienced major sensory losses such as muscle paralysis were still capable of feeling emotions (Hockenbury and Hockenbury, 2010 [43]).

While modern physiologists and psychologists largely discount the James-Lange theory and study it for nothing more than its historical significance, there has been some supporting evidence for parts of the original theory. For instance, studies suggest that our perception of our physiological states is related to the way we experience emotions [11].

Cannon-Bard theory (Bard, 1934 [10])
The Cannon-Bard theory of emotion was first proposed by physiologist Walter Cannon in 1920 and extended by physiologist Philip Bard in 1934 in response to the James-Lange theory, which was the dominant theory at the time. Noting some of the apparent contradictions in

the James-Lange theory, Bard note that we experience both the physiological response and emotions simultaneously.

In the example above, according to the Cannon-Bard theory, the person in question experiences an elevated heart rate increases, and is physiologically aroused due to signals from the autonomic nervous system. At the same time, they also experience the emotion of fear. The two events happen independently.

These theories explain the events leading to some emotions well, but have their shortcomings nevertheless. Modern theories of emotion use our cognition of the environment as the basis for both the physiological and emotional responses.

1.3.2 Cognitive Theories

Cognitive theories of emotion started emerging in the 1960's, and were a part of the "cognitive revolution" in psychology. we discuss two major cognitive theories of emotion here:

Schachter-Singer's Two Factor Theory (Schachter and Singer, 1962 [90])
Also known as the two-factor theory of emotion, the Schachter-Singer Theory is an example of a cognitive theory of emotion. This theory suggests that the physiological arousal occurs first, and then the individual must identify the reason for this arousal to experience and label it as an emotion. In other words, a stimulus leads to a physiological response that is then *cognitively interpreted and labeled* which results in an emotion.

Schachter and Singer's theory draws on both the James-Lange theory and the Cannon-Bard theory of emotion. Like the James-Lange theory, the Schachter-Singer theory proposes that people do infer emotions based on physiological responses. The critical factor is the situation and the cognitive interpretation that people use to label that emotion.

Like the Cannon-Bard theory, the Schachter-Singer theory also suggests that similar physiological responses can produce varying emotions. For example, if you experience a racing heart and sweating palms during an important math exam, you will probably identify the emotion as anxiety. If you experience the same physical responses after a run, you might not associate them to any particular emotion. Another example of differing cognitive interpretations is the situation of being physiologically aroused near a large group of people. If the people are an angry mob, one may label the response as "anger", while the same arousal at a musical concert may be labelled as "excited".

To put their theory to test, Schachter and Singer performed an experiment on 184 male participants by injecting them with epinephrine, a hormone that induces physical arousal

through an elevated heartrate, trembling and rapid breathing. The participants were informed that the drug was to test their eyesight. However, one group was additionally told about the side-effects regarding the physical arousal, while the other group was not.

The participants were then placed in a room with another participant who was actually a confederate of the experiment. The confederate behaved in one of following two ways: euphoric or angry. The experimenters observed that participants who were unaware of the effects of the injection were more likely to feel the same emotion as the confederate than those who were aware. The authors hypothesized that participants who did not have an explanation for their own feelings were more likely to be susceptible to the emotional influence of the confederate.

Other researchers tried to replicate these results, but not all of them with consistent with the original hypothesis, who found that the euphoria in the participants unaware of the side-effects was not significantly different from what they would have had in the presence of a neutral confederate (Marshall and Zimbardo, 1979 [58]). The Schachter-Singer has also received criticism with researchers pointing out that I experience some emotions before thinking about them.

Lazarus's Cognitive Appraisal (Lazarus, 1991 [55])

Over the past few decades, appraisal theory has evolved into a prominent theory in the field of psychology for determining affect and emotion. Acording to this theory, our emotional **and** physiological experience depends on the way I appraise or evaluate the events around us. The various emotions that I experience are simply different types of appraisals of the situations I are in. For instance, while driving down a winding road in the mountains, looking outside the window may result in anxiety for the driver, but may be a calming experience for the passengers.

This theory has its origins in 1945, when Magda Arnold postulated that different emotions such as fear, anger and excitement were a result of different excitatory phenomena (Arnold, 1945 [7]). It was extended by Richard Lazarus, who in 1991 proposed a two-stage structural model for appraisal, comprising of primary and secondary appraisal.

According to Lazarus, the occurrence of an emotion due to a stimulus is always the result of the interaction an individual with the environment. **Primary appraisal** is associated with motivational relevance: an assessment of the individual's goals and how of the circumstances are relevant to attaining them. This aspect controls the intensity of the experienced emotions. **Secondary appraisal** involves the individual's evaluation of whether the resources at their disposal are sufficient for coping with the situation. One aspect of secondary ap-

praisal is the evaluation of *who should be held accountable*: the individual themselves, or another person or group of persons or mere chance. The individual may either blame or credit whoever is accountable for the situation. The way in which an individual views the accountable entity and controls their efforts to cope with their emotions. Another aspect of secondary appraisal is a person's *coping potential*: 1. the ability to make the situation more congruent to one's goals (problem-focused coping) or 2. the ability to handle the situation should the should it remain incongruent to one's goals (emotion-focused coping)(Smith and Kirby, 2009 [92]). The last aspect of secondar appraisal is the *future expectancy* (Lazarus, 1991 [55]), or the individual's expectations of how the motivational congruence of a situation will change in the future. An individual may expect that the situation may change favorably or unfavorably. This determines the emotions resulting from a situation and the coping strategies used.

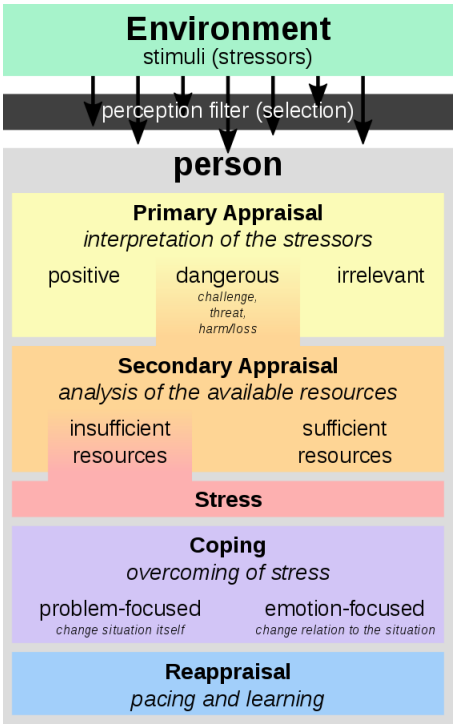


Figure 1.2: Illustration of the Structural Model of Stress and Coping by Richard Lazarus for a stress-inducing situation (source)

As an example, figure 1.2 describes the primary and secondary appraisal of a stress-inducing situation, for instance having to speak in front of an audience. The individual assesses that the situation is "dangerous" in the sense that it can harm their reputation and result in ridicule should they fail. If the individual feels that they haven't prepared well enough (insufficient resources), they will feel stressed. If problem-focused coping fails (they

cannot avoid the situation), then emotion-focused coping takes over and the individual may resort to pacing around to channel their stress.

1.4 MODELLING EMOTIONS

Are Emotions Universal?

There has been significant research on the question of whether there are any core human emotions that exist across all cultures in the world. Darwin believed that facial muscles, which are a universal feature in humans, indicated the emotional state of a person with the aim of enhancing communication, which in turn, increased the chance of survival. In 1971, Ekman and Friesen [33] published a study which supported this viewpoint. They conducted an experiment with the Neolithic, preliterate people in New Guinea, a culture isolated from Western contact until very recently. They instructed a translator to recite well-rehearsed stories, each designed to induce a specific emotion according to Western cultures. After the story, the participants were shown 3 pictures, each displaying a different emotion with only one being correct. Results showed that there was no significant difference between the accuracy achieved by the subjects when compared to the accuracy achieved by literate subjects from Western cultures, which lent strong support towards the belief that at least some emotions are universal.

However, there are cultural influences in the way humans express emotions, as seen in differing interpretations of actions like winking or raising one eyebrow. Moreover, cultures influence the triggers for the display of emotions, for instance, the propensity of Western cultures' to display emotion more openly than their Eastern counterparts. In one study, it was observed that while parents and peers in American societies encouraged the expresion of emotion, Japanese cultures viewed *suppression* of emotion so as to fit with the group more mature and appropriate (Miyamoto et al., 2010 [63]).

Discrete vs Dimensional Models of Emotion

One school of thought is that all humans have an innate set of core emotions that are cross-culturally recognizable. These basic emotions are believed to be *discrete* as they are perceived to be distinguishable by an individual's facial expression and biological processes (Ekman and Friesen, 1971 [33]).

On the other hand, dimensional models of emotion attempt to model emotions by coordinates in two or three dimensional vector spaces. They originate from a belief that there exists a common and interconnected neurophysiological system that is responsible for all emotional states (Posner et al., 2005 [83]). Let us look at these two models in some detail.

1.4.1 Discrete Models of Emotion

Several researchers have proposed ways of organizing emotions into discrete categories, such as Allport (1922) [4], Ekman and Friesen (1971) [33] and Izard (1971) [46]. There have been several theories on which emotions are basic, such as Ekman (1992) [30], Plutchik (1962) [82] and Parrott (2001) [76]. Ekman (1992) [30] argues that there are six basic emotions: joy, sadness, anger, fear, disgust, and surprise.

A leading researcher in the study of human emotion and a supporter of the discrete models, Paul Ekman conducted a survey in 2016 to obtain a concensus on some of the important problems in this domain Ekman (2016) [32]. The participants were active researchers in the field, decided by their publications in journals such as *Journal of Experimental Psychology: General* and the *Proceedings of the National Academy of Sciences*). The results showed that the 88% of the respondents endorsed the view that there was *compelling evidence* for universal features in any aspect of emotion". On the other hand, 55% of respondents believed that both the discrete and the dimensional models were relevant for deciding basic emotions. Coming to the choice of emotions to be considered basic, the major basic emotions endorsed were: anger (91%), fear (90%), disgust (86%), sadness (80%) and happiness (76%). Other emotions such as shame, surprise and sadness were endorsed by 40-50% of the respondents.

However it should be noted that emotions in general do not have clear boundaries and do not always occur in isolation.

1.4.2 Dimensional Emotions

With advances in neuroscience, the constructionist approach has been introduced and studied to analyze the neural basis of emotions from a perspective of broader, less-specific emotional dimensions(Posner et al., 2005 [83]). These dimensional models, as stated before, originate from a belief that there is a common interconnected neurophysiological network that results in all affect states. These are in contrast with theories for basic emotion, which are based on the belief that the specific parts of the brain "produce" different emotions. Dimensional models are most commonly defined, for theoretical as well as practical purposes, according to two to three dimensions. The earliest dimensional model for emotion was proposed by Wilhelm Max Wundt, the father of modern psychology, in 1897. He modeled emotions over three dimensions: 1. pleasurable vs unpleasurable, 2. arousing or subduing and 3. strain or relaxation(Wundt, 1896 [110]). Harold Schlosberg proposed to replace the latter two dimensions by attention–rejection and the level of activation (Schlosberg, 1954 [91]). The popularly used dimensional models today usually incorporate valence (similar

to the pleasurable dimension) and arousal. I describe some of the major models below.

Circumplex Model (Russell, 1980 [86])
Developed by James Russell, the circumplex models emotions in a two-dimensional circular space with arousal and valence dimensions. As shown in figure 1.3, valence represents the horizontal axis and arousal represents the vertical axis. Note that the center of the circle denotes *neutral* valence and a *medium* level of arousal.

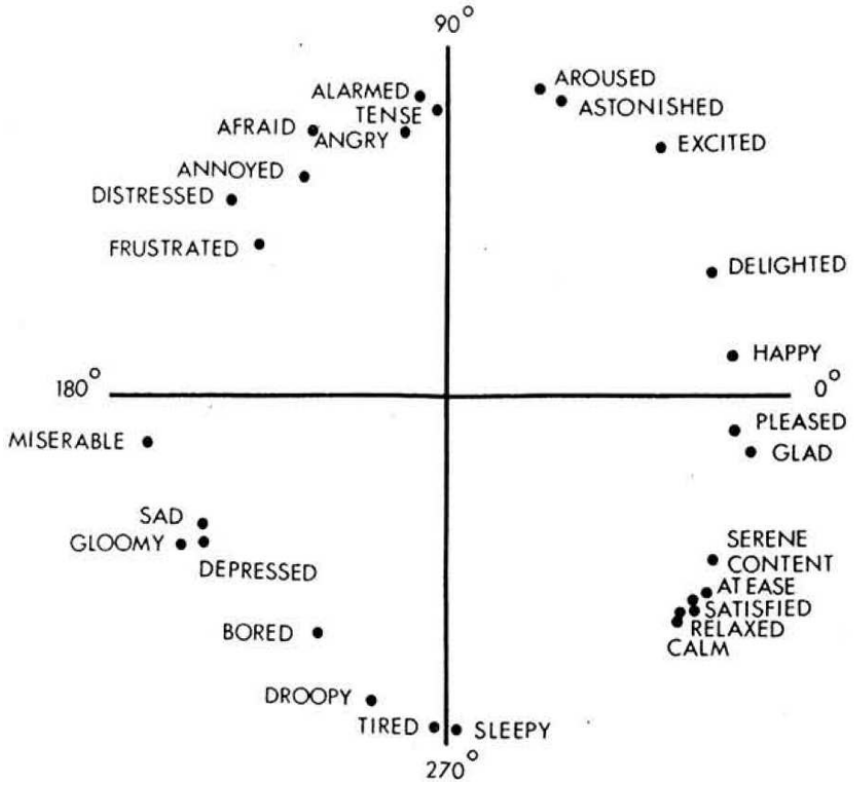


Figure 1.3: Russell's circumplex model (Russell, 1980 [86])

This was later modified to represent core emotions not necessarily directed toward anything (Russell and Barrett, 1999 [87]). The circumplex model has been used for applications such as testing effects of affect-bearing words and facial expressions(Remington et al., 2000 [85]).

PAD Model (Mehrabian, 1980 [59])
PAD uses three numerical dimensions to for representation of all emotions: Pleasure, Arousal and Dominance. The Pleasure-Displeasure dimension quantifies how pleasant an emotion is. For instance both anger and fear score high on the displeasure scale, while joy ranks high on the pleasantness dimension. The Arousal-Nonarousal dimension measures the intensity

of the emotion. For instance, while rage and anger may rank similarly on the pleasantness dimension, rage has a higher arousal value. Tiredness or boredom may correspond to a low arousal value. The Dominance-Submissiveness Scale represents the controlling nature of the emotion. For instance determination and anger are emotions with a high dominance, while fear is a submissive emotion (Mehrabian, 1980 [59]).

Plutchik's Multifactor Theory (Plutchik, 1960 [81])
Plutchik proposed a hybrid theory of emotions that brought together the discrete and dimensional models. His "wheel" of emotions contains eight basic emotions, which act similar to dimensions, including Ekman's six emotions as well as trust and anticipation. This wheel is shown in figure 1.4.

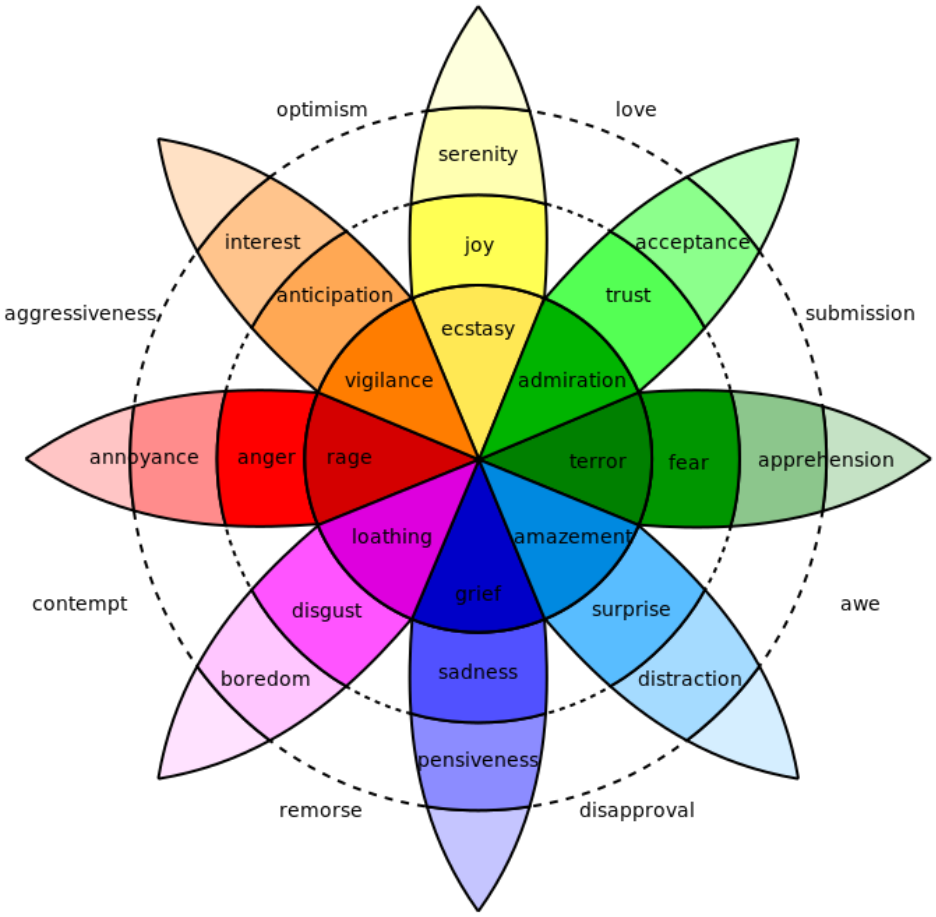


Figure 1.4: Plutchik's wheel of emotion (source)

Parallels to the dimensional approach are drawn by using the radius to indicate intensity

— nearer the emotion is to the center, the higher the intensity. These eight basic emotions form four spatially opposing pairs: joy–sadness, anger–fear, trust–disgust and anticipation–surprise. There are also *dyadic* emotions: emotions that result from a "combination" of two or more emotions. The primary dyads result from two emotions, and are shown in the white space between the basic emotions.

Cowen and Keltner (Cowen and Keltner, 2017 [26]) A recent model by Cowen and Keltner attempts to not only model emotions in a high dimensional space, but also to capture interactions between them through gradients between emotions. Through self-reported emotions over 2185 videos by 9 to 17 subjects, the authors statistically derive a taxonomy of emotion[1]. The authors use the videos to reliably elicit 27 distinct varieties of reported emotional experience. They also make claims for benefits of using the categorical approaches. For instance, categorical labels such as amusement better captured reports of subjective experience than commonly measured affective dimensions like valence and arousal. The authors then modeled the boundaries between emotions by analyzing the gradients of emotion, for instance from anxiety to fear to horror to disgust, or from calmness to aesthetic appreciation to awe. These emotional states occupy a complex, high-dimensional categorical space.

Other models

The **vector model** of emotion (Bradley et al., 1992 [18]) consists of vectors that point in two directions, representing a boomerang shape. It assumes an underlying arousal dimension, and that the valence influences the vertical coordinate of an emotion. High arousal states are differentiated by their valence, whereas low arousal states are more neutral and are represented near the meeting point of the vectors. The **PANA** (Positive Activation - Negative Activation) model or "consensual" model (Watson and Tellegen, 1985 [107]) proposes that positive affect and negative affect originate from two separate systems. Similar to the vector model, states of higher arousal are defined by their valence, and states of lower arousal tend to be more neutral in terms of valence.

1.5 PROBLEMS IN AUTOMATED EMOTION RECOGNITION

While emotion analysis can be applied to all genres of text, certain domains and media tend to contain a stronger presence of emotion-bearing expressions than others. Here, I look at some of the previously explored genres for emotion detection.

[1]Link to interactive map: https://s3-us-west-1.amazonaws.com/emogifs/map.html

News, Blogs

News articles provide an objective description of events over the world. However, it is possible to categorize the emotional content in an article. Strapparava and Mihalcea created a dataset of titles of 1000 news articles extracted from websites such as Google News and CNN and annotated them with the emotional content in them(Strapparava and Mihalcea, 2008 [96]). They created a shared evaluation through a competition in the 2007 iteration of SemEval (Strapparava and Mihalcea, 2007 [95]). The objective was to label each title with the approprate label from a predefined set of emotions (anger, fear, joy, etc.).

Blog posts, on the other hand, are subjective expressions of opinion. There has been previous work on analyzing blog posts for emotion, such as Mihalcea and Liu (2006) [61] and Généreux and Evans (2006) [37].

Social Media Before the days of social media, researchers in affective computing performed emotion recognition on datasets of private communication over the internet. Examples include analysis on emails (Zhe and Boucouvalas, 2002 [112]) and chat messages (Holzman and Pottenger, 2003 [44]). With the rise of social media platforms such as Facebook and Twitter, however, there has been a tremendous growth in opportunity for affect recognition. Tweets are frequently used to convey one's opinion and stance about events. Besides, they are annotated with hashtags that are often representative of the emotional state of the user (#happy, #annoyed, etc), which gives us the opportunity to effectively get pseudo-labelled data using *distant supervision*(Mohammad and Bravo-Marquez, 2017 [68]).

Fiction Literary texts such as novels and fairy tales contain a rich and artistic expression of emotions. Thanks to digitization of texts through platforms like Project Gutenburg[2], I have convenient access to large volumes of literary texts. Project Gutenberg provides access to over 56,000 e-books as of April 2018. There has been work on annotating such work on a sentence level with the emotion content. Alm et al. (2005) [5] annotated a corpus of approximately 185 children stories, by Grimms, H. C. Andersen and B. Potter.

Visualization of Emotions There has also been some interesting work in visualizing emotions, for example that of Subasic and Huettner (2001) [98] and Kalra and Karahalios (2005) [47]. Mohammad (2011) [64] used a large word lexicon, Emolex (Mohammad and Turney, 2013 [70]), to compute a bag-of-words score for the emotion of a sentence. They then tracked the progression of emotions in novels.

[2]Website: http://www.gutenberg.org/

Empirical assessment of emotions in literary texts has sometimes relied on human annotation of the texts, but this has restricted the number of texts an- alyzed. For example, Alm and Sproat (2005) an- notated 22 Brothers Grimm fairy tales to show that fairy tales often began with a neutral sentence and ended with a happy sentence.

1.6 CHALLENGES IN AUTOMATED EMOTION RECOGNITION

Cost of Annotating Data

One major challenge in creating an annotated resource with emotion labels is the high cost and considerable human effort involved in the process. Lately, however, with the evolution of social media and crowdsourcing, this problem has been mitigated to some extent. News websites and social media platform allow users to express their emotional reactions to posts and articles, which is potentially a powerful source of annotations without any need for expertise. Crowdsourcing (Howe, 2006 [45]), the idea of breaking down a task into small independent subtasks and distributing them to a large number of people (usually over the web), has also opened new avenues for inexpensive creation of large emotion labeled corpora. Platforms like Mechanical Turk are very commonly used to perform this task along with many others. However, the first and foremost requirement in such settings is always quality control. The task and compensation may attract cheaters who input random responses and malicious annotators who input incorrect information. The onus is on the researchers who create datasets to provide concise and easily understandable instructions that can be understood by the general users of crowdsourcing platforms (Mohammad and Turney, 2013 [70]).

No context for short text

Platforms such as Twitter have a large and diverse user base which in turn results in rich textual content such as the use of colloquial and sometimes non-standard language, such as emoticons, creative spellings (“wut”, “noooo”, “happee”, etc.) and hashtags (#picftheday, #like4like, etc.). However, the length of the text can also pose some challenges. The emotion evoked by a word depends on the context. For example, the emotion evoked by the word shout is different in the context of admonishment than when used in the following context: “Give me a shout if you need any help.” (Ghosh et al., 2015 [39]).

Figurative Language

Creative usage of text may make it difficult for automated systems to infer the emotion. In the case of figurative devices like irony, sarcasm and metaphor, often, secondary or extended

meaning, rather than the literal meaning of words is the intent of the user. So significant are these devices, that they can even invert the polarity of a sentence. "Yeah, right" can fool any bag of words model with the presence of two affirmative words in it. When viewed sarcastically, however ("Yeah, right #sarcasm"), I understand it expresses a negative sentiment. Other devices like irony may also contain affirmative language to convey critical meanings. Figurative language tests the limits of traditional methods for supposedly literal texts. This problem is important since figurative language is pervasive in almost any genre of text, and is especially used in social media.

In this chapter, I have explored the various theories of emotions. I studied discrete and dimensional approaches for modeling the emotional state of an individual. I also looked at some applications for emotion detection in text and some challenges that I face while working with different genres of text, such as fiction, news and short texts on social media. The layout of this thesis is as follows. Chapter 2 explores some of the past work in the area of affective computing in text. Chapter 3 details the approaches that I worked with. Chapter 4 contains experimental details and results, and chapter 5 addresses some of my future work in this regard.

Chapter 2 Related Work

In this chapter, I look at some of the related work in affect recognition in text. Most approaches in this regard are supervised. I divide this chapter as follows. I explore word-emotion lexicons in the literature based on whether they model emotions as discrete entities or continuous entities in a vector space. I also look at word embeddings, which are increasingly being used in recent models. I then study some of the approaches that have been proposed to infer the emotion using these features. Finally, I enumerate some shared evaluations such as SemEval that enable the creation of benchmarks for models based on a common evaluation metric and dataset.

2.1 WORD-EMOTION LEXICONS

Word-emotion lexicons are a mapping between the words in the vocabulary to an emotion rating. Emotions modeled can be either discrete or dimensional, as we saw in chapter 1. Let us look at some popularly used lexicons in the literature.

2.1.1 Discrete Emotions

As we saw in chapter 1, there are several taxonomies for modeling emotions as discrete entities. Several lexicons use a set of emotions similar to the set of Ekman's basic emotions [30], which divides all emotion into six basic categories: joy, sadness, fear, disgust, anger and surprise.

1. **General Inquirer** (Stone et al., 1962 [94])
   General Inquirer (GI) has about 12,000 words labeled with 182 tags. Some of these tags are indicative of valence, such as the positive and negative tags. Some tags describe other affect categories such as pleasure, arousal, feeling, and pain.

2. **Wordnet Affect**(Strapparava et al., 2004 [97])
   Created in 2004, Wordnet Affect contains a subset of synsets from the larger Wordnet Domains lexicon for affective analysis of text. It contains a hierarchical tagging of emotions. Emotions are clustered into positive, negative, ambiguous (context dependent) or neutral. Elements deeper in the hierarchy represent more fine-grained differences in emotions. The entire hierarchy comprises of over 300 nodes. The lexicon links synsets to each of the leaf nodes in the hierarchy. It contains about 900 annotated synsets and 1.6k words annotated as (lemma, POS tag, sense) triplets.

3. **NRC-10 Emotion Lexicon** (Mohammad and Turney, 2013 [70])
   Also called the NRC Emotion Lexicon, this corpus, crowdsourced using the Mechanical Turk, is an order of magnitude larger than the Wordnet Affect corpus, and contains binary tags corresponding to ten emotion labels (joy, sadness, anger, surprise, fear, disgust, trust and anticipation; and two sentiment classes: positive and negative) for over 14,000 words. A word may be tagged with multiple emotions. It provides tags for direct as well as indirect affective words (example: fearful, a direct affective word and monster, an indirect affective word). One must be careful, however, about the ratings in this corpus, as they are taken without context.

4. **NRC-10 Expanded Lexicon** (Bravo-Marquez et al., 2016 [19])
   The NRC-10 lexicon, being hand-labeled, is limited in its coverage. Bravo-Marquez et al. expanded the NRC-10 lexicon using an unlabelled corpus of 10 million tweets. First, the NRC-10 words frequent in the unlabelled corpus ($\geq$ 50 occurrences) were identified. A word2vec model was trained on these tweets, and a multi-class classifier was trained on the word-level embeddings using the original NRC-10 words as ground truth. The NRC-10 expanded lexicon contains about 43,000 additional words with respect to the NRC-10 lexicon, each with 10 emotion labels.

5. **The Fuzzy Affect Lexicon** (Subasic and Huettner, 2001 [98])
   This lexicon contains roughly 4,000 `lemma#PoS` manually annotated by one linguist using 80 emotion labels.

6. **Sentiment Lexicons**
   There are several lexicons that contain word-level sentiment association scores. I enumerate a few of them below.

   - MPQA (Wilson et al., 2005 [109])
   - BingLiu (Bauman et al., 2017 [13])
   - AFINN (Nielsen, 2011 [73])
   - SentiWordNet (Baccianella et al., 2010 [8])
   - Sentiment-140 Emoticons (Kiritchenko et al., 2014 [51])
   - SentiFul (Neviarouskaya et al., 2009 [72]) assigns sentiment features to words, such as sentiment features: propagating, reversing, intensifying, and weakening and is automatically expanded over a set of seed words using direct synonymy, antonymy, hyponymy, derivation, and compounding relations.

2.1.2 Continuous Emotions

An alternate approach for classifying emotions is to treat them as elements of a vector space. In this regard, the most common representation of emotions is in a three dimensional space whose axes correspond to the valence, arousal and dominance (VAD) felt by the subject of the emotion. Some word-emotion lexicons that model the emotions in a continuous space are:

1. **Affective Norms for English Words (ANEW)** [17]: This corpus contains VAD ratings for about 600 words. Each subject responsible for annotation assigns a discrete score for the valence, arousal and dominance of each word along a 9-point rating scale. The final scores are averaged over all subjects.

2. **WKB Corpus** [106]: This corpus is an extension of the ANEW corpus. It contains VAD annotations for about 14,000 English lemmas performed by participants on Amazon's Mechanical Turk website. Similar to the ANEW corpus, the individual ratings are on a 9-point scale, and the final ratings are averaged over the individual ratings.

3. **NRC Hashtag Emotion Lexicon** (Mohammad and Kiritchenko, 2015 [69])
   This corpus contains a real valued mapping between hashtags and their emotion association. The corpus was semi-automatically generated. The authors retrieved about 21,000 tweets with hashtags pertaining to Ekman's six emotions (#fear, #anger, etc.). They then computed the strength of association (a metric inspired by the pointwise mutual information) between n-grams and emotions for about 11,400 n-grams.

4. **DepecheMood** (Staiano and Guerini, 2014 [93])
   Similar to the NRC-10 corpus, DepecheMood contains 37,000 terms automatically annotated with emotion scores. The dataset used contained about 25,300 articles from `rappler.com` annotated with emotion scores. The emotion scores are obtained by performing transformations on the word-document and the document-emotion matrices constructed using news articles as documents. Each word has fractional scores for emotions (afraid, amused, angry, annoyed, don't care, happy, inspired, sad) it conveys.

5. **SenticNet 5**(Cambria et al., 2018 [23])
   SenticNet 5 contains conceptual primitives extracted from text and stored as concept embeddings. These are linked to commonsense concepts and named entities in a new three-level knowledge representation.

2.2 SENTENCE LEVEL LABELED CORPORA

Large scale corpora annotated with sentence-level emotion labels are uncommon in the literature. In this section, I enumerate to the best of my knowledge the major datasets that contain emotion labels for phrases or sentences.

1. **Affective Text** (Strapparava and Mihalcea, 2007 [95])
   This corpus, created for SemEval 2007 (Strapparava and Mihalcea, 2007 [95]) to perform affective analysis of news articles, contains 1000 test headlines and 200 development headlines, each annotated by 6 annotators. The space of annotations spans the Ekman labels and the valence labels (positive or negative).

2. **Annotated Children's Stories** (Alm et al., 2005 [5])
   This corpus contains approximately 185 children stories, by Grimms, H. C. Andersen and B. Potter, annotated at a sentence level by at least 2 annotators for the six Ekman Labels. The dataset has high inter-annotator agreement, and also provides POS tagged and preprocessed data.

3. **Emotional Phrase and Sentence annotated data** (Aman and Szpakowicz, 2007 [6])
   This dataset contains 5,000 sentences each annotated with emotions spanning 8 categories: Ekman's emotions along with the neutral, mixed categories. Each sentence also contains an annotation for the intensity and emotion bearing phrases. There are two annotators per sentence. To gather the dataset, the authors use seed words and retrieve all blogs with those words. They then analyze each sentence from those blogs.

4. **Emotex** (Hasan et al., 2014 [40])
   Since it is difficult to manually annotate the emotions present in sentences, the authors use twitter hashtags that are indicative of certain emotions. They compare the goodness of the hashtags to classify emotions, and demonstrate positive results for the same.

5. **Valence and Arousal on Facebook Posts** (Preotiuc-Pietro et al., 2016 [84])
   This dataset contains about 3,000 social media posts annotated by two psychologically trained annotators for the valence and arousal, making this one of the few datasets that contains annotations based on the VAD model.

6. **Datasets on CrowdFlower**: There are a few datasets on CrowdFlower that are relevant to emotional analysis of text. This dataset 18 emotional annotations based on Plutchik's wheel of emotions. This dataset, on the other hand, contains emotion annotations for about 40,000 tweets.

2.3 APPROACHES

2.3.1 Rule Based Approaches

Rule based features incorporate domain knowledge. This can include term-based n-gram features. Some methods rely on distance between certain terms. Selected phrases chosen by pre-specified POS patterns, usually including an adjective or adverb, have also been used to perform emotion detection. Early work in this area focused mainly on linguistic heuristics. For example, in their work on sentiment detection, Hatzivassiloglou and McKeown (1997) [41] discuss how two classes of interest give rise to opposite constraints that help the system label decisions. Turney (2002) [101] classified items based on fixed phrases for expression of opinions. Kamps et al. (2001) [48] and his colleagues classified items by bootstrapping, using a seed set of opinion words and a knowledge base like WordNet. Kim and Hovy (2004) [50] used semantic frames to identify sentimental topics. However, a major drawback of these rule-based approaches is that they are unable to detect novel expression of sentiment.

2.3.2 Keyword-Based Approaches

Keyword based approaches classify text based on the detection of unambiguous words in language(Wiebe et al., 2005 [108]). They depend on large scale lexicons with affective labels for words, such NRC (Mohammad and Turney, 2013 [70]) and General Inquirer (Stone et al., 1962 [94]). However, given that they are shallow word level classifiers, they have some significant weaknesses: for instance, they can't reliably recognize affect-negated phrases. For instance, a keyword based approach is likely to correctly classify the sentence "You did a wonderful job." as being of positive valence, but is also likely to assign the same classification to "You did not a wonderful job.". They do not work on sentences that contain strong emotions without the usage of explicit emotion-bearing words. For instance, the sentence: "My brother underwent surgery and will have to be bed-ridden for over six months." evokes a strong negative emotions without using affect keywords, and therefore a keyword-based approach in this case is likely to be ineffective.

2.3.3 Knowledge-Based Approaches

These methods use web ontologies or semantic networks to affective analysis. A major advantage of such systems is that they enable the system to use conceptual ideas derived from world knowledge. By relying on large semantic knowledge bases, such approaches overcome

the shallow nature of keyword based analysis and word co-occurrence counts, and instead rely on the implicit meaning/features associated with natural language concepts. Concept-based approaches analyze multi-word expressions that don't bear emotion explicitly, but that point to concepts that do. Their fixed/flat representa-tion, finally, places bounds on inferences of semantic and affective fea-tures associated with concepts. **Sentic Computing** approaches this problem using a three-layer structure that performs tasks like multiword concept extraction, named-entity recognition, discovery of conceptual primitives from text and links to commonsense concepts (Cambria and Hussain, 2012 [22]).

2.3.4 Distributed Approaches

Recently, distributed approaches have been proposed to produce effective word representations. For example, in the Word2Vec model (Mikolov et al., 2013 [62]), by maximising the likelihood of observing a word within a fixed context window, it is possible to learn for each word in the vocabulary a dense real-valued embedding vector from even a shallow neural network. A consequence of this representation is that similar words are close to each other in the embedding space. The popularly used GloVe model (Pennington et al., 2014 [79]) aims to capture global corpus statistics through the word co-occurrence probabilities. For the purpose of document classification, it is important to design a function that maps a sequence of embedding vectors to a document-embedding vector. Averaging the word embeddings is a popular strategy (Mohammad and Bravo-Marquez, 2017 [68]). However, more general learning strategies such as sequence-to-sequence learning using recurrent neural networks (Sutskever et al., 2014 [99]) and document encoding using convolutional neural networks (Wang et al., 2012 [105]) is usually preferred. These methods have been applied to specific domains such as Twiiter (Vosoughi et al., 2016 [104]).

2.3.5 Statistical Approaches

These approaches use sparse word-level features such as n-grams and apply statistical methods such as Bayesian inference and support vector machines on them for classification. Early examples include work by Pang et al. (2002) [74] on movie review classification and many others. The appeal of such systems lies in the possibility that by feeding a machine-learning algorithm a large training corpus of annotated texts, the system may not only learn the affective valence of affect keywords (like in the keyword-spotting approach), but also take into account the valence of other arbitrary keywords that are typically not considered in keyword based approaches along with other factors such as punctuation and

word co-occurrence frequencies. Recent work on statistical approaches has utilized more dense representations as described in subsection 2.3.4. Favoured architectures employ neural networks over these dense word embeddings to transform the input document into the embedding space, followed by shallow learners that can predict the class or the intensity of emotion in the document(Mohammad and Bravo-Marquez, 2017 [66]). However, statistical methods generally tend to work well only when they receive sufficiently large text input.

2.4 SHARED EVALUATIONS

Shared evaluations enable a community to establish a benchmark for the task-specific systems through the use of a common dataset and evaluation metric. This expedites the discussion of what techniques work and what don't, along with an analysis of them. Over the past few years, there has been an increase in the organization and participation in shared evaluations for Affect Recognition in text, especially in Tweets. This is largely due to the potential for performing distant supervision over Twitter, due to the tagging of tweets with emotion-bearing hashtags. I enumerate some of the major evaluations here.

1. **Affective Text**, SemEval 2007 (Strapparava and Mihalcea, 2007 [95])
   The earliest instance of a shared evaluation in this domain, this task focused on the classification of emotions and valence in news headlines, and was meant as an exploration of the connection between emotions and lexical semantics.

2. **Shared Task on Emotion Intensity**, WASSA 2017 (Mohammad and Bravo-Marquez, 2017 [66])
   This task focused on computing the intensity of emotion felt by the speaker of a tweet. The organizers created the first datasets of tweets annotated for anger, fear, joy and sadness intensities using a technique called best–worst scaling (BWS). Systems were evaluated using the Pearson evaluation between their predictions and the annotations. The best performing teams used a combination of lexicon-based features and CNN-LSTM based document representations using word embeddings (Mohammad and Bravo-Marquez, 2017 [66]).

3. **Affective Tweets**, SemEval 2018 (Mohammad et al., 2018 [71])
   This task was an extended version of the Shared Task on Emotion Intensity at WASSA 2017. The organizers provided a dataset constructed using BWS. There were various sub-tasks, such predicting the emotion intensity, assigning tweets interpretable scores

(-3 to +3) for each emotion and valence detection. These tasks were organized for English, Spanish and Arabic tweets.

Chapter 3 Approaches

In this chapter, I will describe the theory behind the various features and models that I followed building our emotion prediction system. I first begin by describing the features used in our system. Specifically, I look at both traditional lexicon-based features and the more recent embedding-based features. Then, I describe the motivation and theory behind the models that I use, ranging from the shallow linear models to tree-based models to non-linear deep neural network based models.

3.1 FEATURES

In this section, I go into the details of the traditional lexicon-based features and the distributed embedding based features.

3.1.1 Lexicon-Based Features

As I saw in section 2.1, lexical features, obtained from word-emotion lexicons, contain shallow yet significantly useful information about the affective content in text. The shallowness in depth results in lexicons suffering from the issue of having low recall. To mitigate the issue to some extent, I extract features from several word-lexicons, some of which provide word-emotion features and others, word-sentiment features. In particular, I use the following word-affect lexicons to extract the lexical features:

- MPQA (Wilson et al., 2005 [109])
- BingLiu (Bauman et al., 2017 [13])
- AFINN (Nielsen, 2011 [73])
- NRC Hashtag Emotion Lexicon (Mohammad and Kiritchenko, 2015 [69])
- SentiWordNet (Baccianella et al., 2010 [8])
- Sentiment-140 Emoticons (Kiritchenko et al., 2014 [51])
- NRC emotion lexicon (wordlevel) (Mohammad and Turney, 2013 [70])
- NRC-10 Expanded Lexicon (Bravo-Marquez et al., 2016 [19])
- NRC-AffectIntensity-Lexicon

- SentiWordNet (Baccianella et al., 2010 [8])
- NegatingWordList (Mohammad and Bravo-Marquez, 2017 [67])

The above lexicons were procured from the AffectiveTweets package (Mohammad and Bravo-Marquez, 2017 [67]) that has been implemented using Weka.

3.1.2 Word Embeddings

As I saw in section 2.3.4, distributed approaches have become popular in recent years to produce effective word representations. Some popularly used word embeddings in the literature are the Word2Vec embedding (Mikolov et al., 2013 [62]), and the GloVe model (Pennington et al., 2014 [79]). For niche tasks like Twitter Emotion Detection, I use word embeddings that have been specifically trained on tweets. In our experiments, I use word embeddings trained on the Edinburgh corpus (Petrović et al., 2010 [80]), that contrains about 97 million tweets. I experiment with both the 100- and 400-dimensional Word2Vec embeddings .

3.2 MODEL

In this section, I describe the models used in our system. I begin with the baseline - shallow linear models - followed by tree-based models and deep neural network-based models.

3.2.1 Linear Models

Linear models are standard learners in machine learning settings, in the capacity of classifiers and regressors. I look at some popularly used linear learners in the literature in this subsection. Here, I assume the following notation: $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix containing the training data (typically $n \gg p$) and $\mathbf{y} \in \mathbb{R}^n$ is vector containing the labels. $\mathbf{w} \in \mathbb{R}^p$ is the vector of linear weights. Linear models model the output as:

$$y = f(\mathbf{w} \cdot \mathbf{x}) \tag{3.1}$$

where $\mathbf{x}, y$ are individual data point and label respectively. In linear classifiers, y takes on a finite number of discrete real values.

- Linear Regression
  I compute the linear weights $\mathbf{w}$ as the minimizer to the following least squares problem:

$$\mathbf{Xw} \approx \mathbf{y} \tag{3.2}$$

This model is typically adapted to the classification problem as a logistic regressor. In the case of a K-class classification setup, I learn K linear decision boundaries and make the decision based on the softmax function, which chooses the class that maximizes:

$$\frac{e^{\mathbf{w}_i \cdot \mathbf{x}}}{\sum_j e^{\mathbf{w}_j \cdot \mathbf{x}}} \tag{3.3}$$

- Support Vector Machines (SVM)
  SVM's are a representation of the data as points in space, mapped so that the examples of different categories are divided by a clear gap that is as wide as possible (the maximum margin hyperplane). They can also be used as linear regressors. In addition to performing linear classification and regression, SVMs can efficiently perform a non-linear classification the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. For linearly non-separable data points, the SVM is trained using the hinge loss function:

$$\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i)\right) + \lambda ||\mathbf{w}||_2^2 \tag{3.4}$$

  where $\mathbf{x}_i, y_i$ are an individual data point and label respectively, and λ is a user defined parameter. Smaller value of λ, the closer the classifier is to the hard classifier for linearly separable points.

While easy to implement, linear models have their pitfalls. One must be careful with linear models, as they may be affected by outliers in the data. They work poorly with data that do not have a linear relationship between their variables. I experiment with several of these linear models and establish a strong baseline by using them in conjunction with lexical features.

3.2.2 Tree Based Models

Decision Trees

Decision trees are predictive models that make splits over the features in data sequentially to make decisions on class labels. In classification tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels (Safavian and

Landgrebe, 1991 [88]). In regression tree structures, the target variable can take continuous values (typically real numbers). Splits are made based on some measure of "impurity" (for instance, Gini index) based on the distribution of data points under a node.

Decision trees are advantageous when features have associated meanings, leading to interpretability in the model's decisions. Also, decision trees are invariant to scale of the data, thereby eliminating the need for normalization. They can handle both numerical and categorical data. Most implementations of decision trees are variations of the original CART algorithm (Breiman, 2017 [21]).

However, decision trees have certain pitfalls. They may not be robust - small changes to the data may result in large changes to the tree structure. The problem of learning the optimal tree is NP-complete due to the exponential growth in the space of feature splits with the number of features. Algorithms such as CART, which perform greedy splits, may not produce the optimal tree structure. Also, decision trees may overfit the data and as a result, may not generalize well.

Random Forests

Random forests (Ho, 1995 [42]) are essentially ensemble of shallow decision trees. They operate by constructing several potentially shallow decision trees at training time and providing the mode class (for classification) or mean prediction (for regression) of the individual trees. Random forests correct for the shortcoming of decision trees that results in overfitting to the training data.

Gradient Boosted Trees Gradient boosting, first introduced by Breiman (1997) [20], is a general technique for which produces a prediction model in the form of an ensemble of weak prediction models. The model is constructed in a stage-wise manner, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. These trees are sequentially trained to minimize the error resulting from the cumulative set of trees constructed previously. Implementations such as XGBoost (Chen and Guestrin, 2016 [24]) and LightGBM (Ke et al., 2017 [49]) are popular for predictive modeling, and are often used as powerful baselines on which bigger models are built.

While the gradient boosted trees work well for typical predictive modeling tasks, one must be careful with the choice of variables. Gradient boosting with high depth captures interaction between variables well, but may degrade the model's generalizability. Regularization by penalizing complex trees and subsampling the data for fitting restricts the predictive power of the tree, and may mitigate the overfitting.

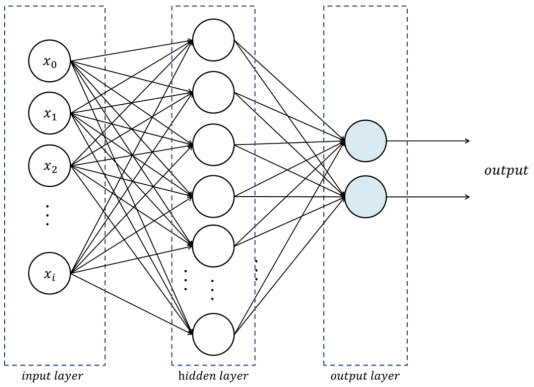


Figure 3.1: A feedforward neural network (source)

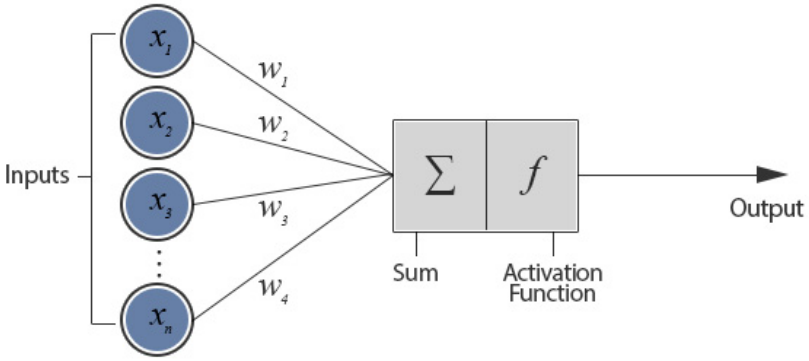


Figure 3.2: An individual neuron (source)

3.2.3 Feedforward Neural Networks

A feedforward neural network (Bebis and Georgiopoulos, 1994 [14]) is an *acyclic* artificial neural network architecture. One of the most popularly used class of neural networks used by the deep learning community is the multilayer perceptron. It consists of multiple layers of computational units, usually interconnected in a feed-forward way. Figure 3.1 shows a typical feedforward architecture.

Each neuron (shown in figure 3.2) in one layer has weighted directed connections to the neurons of the subsequent layer. Neurons have activation functions that transform the feature space. Often, these activation functions introduce non-linearity to the manifold in which the features lie. Popular choices for activation functions include the sigmoid and the ReLU functions. In many applications the units of these networks apply a sigmoid function as an activation function.

Multilayer perceptrons are typically trained using back-propagation. In each iteration, the gradient of the predefined error function is computed with respect to all the parameters

(or weights) in the network, and the parameters are updated using convex optimization (and typically gradient based) methods. To enable learning over large datasets, weights are typically updated after processing small batches of data. Methods like momentum are used to stabilize the learning process.

Multi-layer perceptron, feedforward neural networks, convoluted neural netrowks (CNN), long short term memory networks (LSTM). Neural networks capture non-linear relationships between the features through a series of non-linear transformations on the manifold that is the feature space. Hyperparameter optimization is necessary with neural networks, as is using the appropriate non-linearity in each layer and the correct learning algorithm, which is typically gradient descent with some form of momentum.

3.2.4 Sequence Learning with Neural Architectures

LSTM

Recurrent neural networks are a class of neural networks where connections between units form a directed graph along a sequence, enabling it to exhibit *dynamic temporal behavior* for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state to process input sequences. The RNN equation is given by:

$$h_t = \phi(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1}) \tag{3.5}$$

$$y_t = \mathbf{V}\mathbf{h}_t \tag{3.6}$$

where $\mathbf{U}, \mathbf{W}, \mathbf{V}$ are transition matrices for the input, state-to-state and state-to-output transitions respectively. Training an RNN involves learning these transition matrices. Training RNN's can be difficult for sequences with long-range dependencies (Pascanu et al., 2013 [77]). LSTM's (Gers et al., 1999 [38]) mitigate this problem by the introduction of gates. Figure 3.3 shows the LSTM architecture unrolled over time. The equations governing the LSTM cell are shown in figure 3.4. LSTM's have been very effective in several tasks, and have set state-of-the-art benchmarks in problems such as speech recognition (Sak et al., 2014 [89]), machine translation (Sutskever et al., 2014 [99]) and image captioning (Vinyals et al., 2015 [103]).

CNN

Convolutional neural netoworksa are a class of deep, feed-forward neural networks that are commonly used in computer vision. CNNs use relatively little pre-processing compared to traditional image classification algorithms. They learn the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human

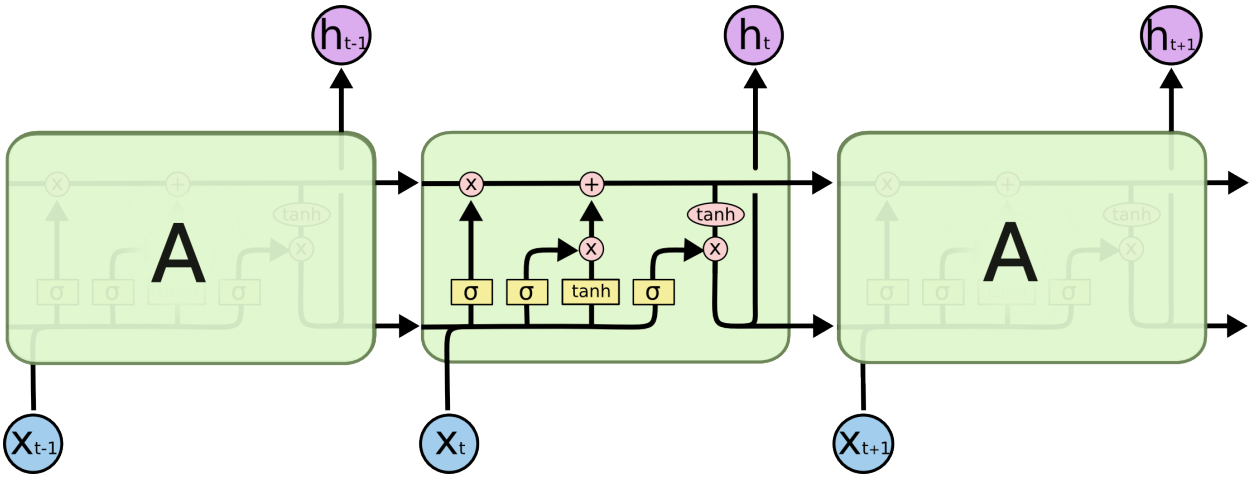
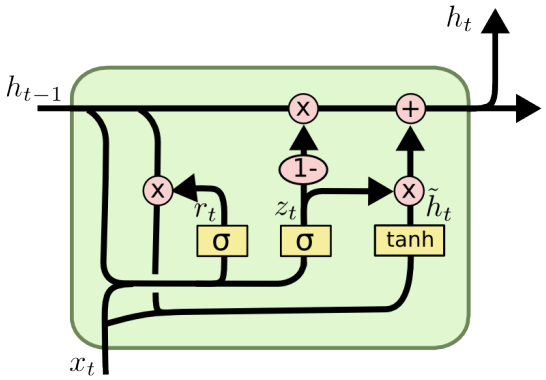


Figure 3.3: A feedforward neural network (source)



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 3.4: An individual neuron (source)

effort in feature design is a major advantage of CNNs. Recently, CNN's have also been explored for text classification (Wang et al., 2012 [105]), as seen in figure 3.5.

Incorporating Attention

Attention mechanisms in neural networks are a recent development. Very loosely based on the visual attention mechanism found in humans, attention mechanisms "allow" the network to weigh specific parts of the input more than others while making predictions. Seen another way, attention mechanisms simply give the network access to its internal memory (Bahdanau et al., 2014 [9]). In our implementation, I use the attention mechanism proposed by Tan et al. (2015) [100] for question answering.

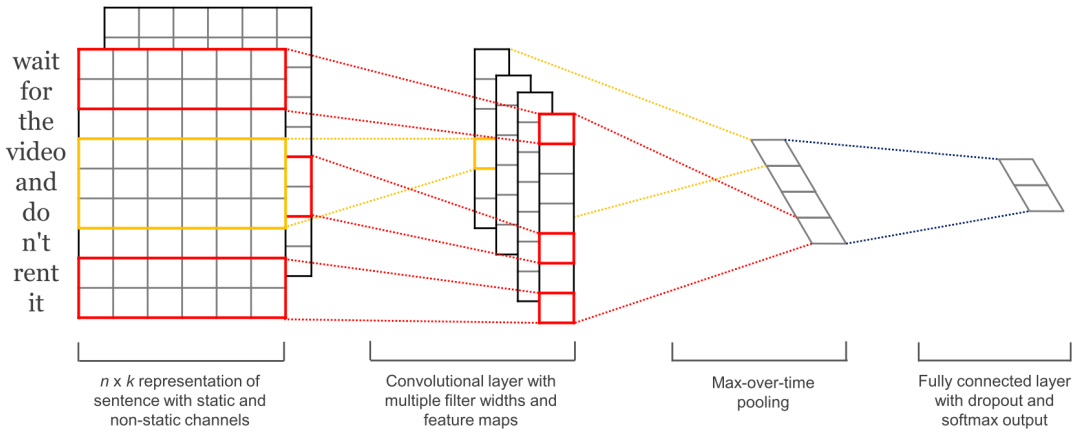


Figure 3.5: A convolutional neural network used for text classification (source: Zhang and Wallace (2015) [111])

3.3 CONSTRUCTING AN ENSEMBLE

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods seek to construct a set of hypotheses and use them in combination.

Ensemble methods [29] have been very popular over the last few years in predictive modeling competitions, with the winning model in various data science competitions almost always being an ensemble. There are several variants of ensemble methods, such as bagging, boosting and stacking, to name a few.

The several learners contained in an ensemble are usually called *base learners*. The generalization capability of an ensemble is usually significantly higher than the individual base learners. Part of the reason why ensemble methods are appealing is that they are able to combine the capability of several *weak learners* and generate strong learners that can make accurate predictions.

One approach to ensembling is **stacking**. During stacking, one concatenates the predictions from all base learners and trains a model to fit the ground truth data using the *predictions of the individual learners* as features. *Feedforward stacked layers* are a powerful extension of this idea. Any one layer of a typical ensemble uses training data in the form of features and test data in the form of true labels. For layers beyond layer 1, the training data provided to them are *predictions from previous layers*. While there are tools such as StackNet [60] that can stack models in layers, I have implemented our own stacked layer

models. The working of a such layer (say, layer N) is as follows:

I first split up the train data into sets of features (in our case, lexicon-based features and embedding-based features). I then train multiple classifiers on each of the multiple feature sets. I obtain out-of-fold predictions for the training set through cross validation. These predictions are now the training set for layer 2. It is possible to generalize this idea to more than 2 layers, but I perform our experiments with a 2-layer ensemble.

Chapter 4 Experiments and Results

In this chapter, I describe the details of SemEval 2018, which is the basis for our experiments. I also describe the model I develop in detail and show its performance across various subtasks in SemEval 2018. Finally, I perform an error analysis of our model over selected sentences and use that as the basis for potential improvements to the model.

4.1 SEMEVAL 2018

Task 1 in SemEval 2018 contains an array of tasks where systems have to automatically determine the intensity of emotions (referred to as "E") and intensity of sentiment (i.e. the valence "V") of the tweeters (the poster of the tweet) from their tweets. The authors provide a separate training and test datasets for English, Arabic, and Spanish tweets for all substasks. The individual tasks are described below:

1. **EI-reg** (an emotion intensity regression task): Given a tweet and an emotion E, determine the intensity of E that best represents the mental state of the tweeter—a real-valued score between 0 (least E) and 1 (most E). Separate datasets are provided for anger, fear, joy, and sadness.

2. **EI-oc** (an emotion intensity ordinal classification task): Given a tweet and an emotion E, classify the tweet into one of four ordinal classes of intensity of E that best represents the mental state of the tweeter. Separate datasets are provided for anger, fear, joy, and sadness.

3. **V-reg** (a sentiment intensity regression task): Given a tweet, determine the intensity of sentiment or valence (V) that best represents the mental state of the tweeter—a real-valued score between 0 (most negative) and 1 (most positive).

4. **V-oc** (a sentiment analysis, ordinal classification, task): Given a tweet, classify it into one of seven ordinal classes, corresponding to various levels of positive and negative sentiment intensity, that best represents the mental state of the tweeter.

5. **E-c** (an emotion classification task): Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions that best represent the mental state of the tweeter.

Here, E refers to emotion, EI refers to emotion intensity, V refers to valence or sentiment intensity, reg refers to regression, oc refers to ordinal classification, c refers to classification.

The tasks are summarized in table 4.1. I test our models for **subtasks 1-4** trained on the English language tweets in the dataset.

ID	Task Label	Input	Output
1	El-reg	Tweet (t), Emotion (e)	Intensity(e, t) $\in$ (0, 1)
2	El-oc	Tweet (t), Emotion (e)	$0 \leq$ Intensity(e, t) ≤ 3, Intensity(e, t) $\in \mathbf{N}$
3	V-reg	Tweet (t), Sentiment (s)	Intensity(s, t) $\in$ (0, 1)
4	V-oc	Tweet (t)	-3 $\leq$ Intensity(s, t) ≤ 3, Intensity(s, t) $\in \mathbf{N}$
5	E-c	Tweet (t)	Class: neutral/ no emotion/ multiple emotions

Table 4.1: Description of the five sub-tasks of Task1: Affect in Tweets at SemEval 2018.

Dataset Details

Subtasks 1 and 2 (emotion intensity and emotion ordinal classification) share the same training and development data sets: a total of 7,500 sentences in training and about 1,600 sentences in development across the four emotions: anger, fear, joy, sadness. It is interesting to note that the training data sets for the emotions of fear, anger and sadness overlap significantly: all pairs have a Jaccard similarity of over 0.5. This means that over 67% of the data sets across these emotions contain the same tweets.

Subtasks 3 and 4 (valence regression and valence ordinal classification) share the same data sets as well, for a total of 1,200 tweets in training and 450 tweets in development across the four emotions.

Another interesting overlap is between the tweet collections for subtask 5 (emotion classification) and subtask 1: The data set for subtask 5 is made up largely of the tweets for subtask 1, for both the training and development sets. These overlaps of the training and development data sets across all emotions gave us the idea to tackle all tasks using a common set of features. For instance, subtasks 2 and 4 may be solved by simply transforming the output of subtasks 1 and 3, respectively. Task 5 involves a multi-label classification and thus, needs more thought.

In the test set, with the exception of the first 1,000 or so sentences, nearly 95% of the total sentences for subtasks 1A and 3A (i.e., for English) are the so called "mystery" sentences – meaning, essentially neutral sentences without any emotional content. I report the results

on the non-mystery subset of the test dataset.

4.2 PIPELINE

Preprocessing

I first begin by preprocessing the dataset. To extract the lexical features and embedding-based features from the data, I remove all punctuations and convert the text to lower case. I lemmatize the text and use the Weka package *AffectiveTweets* (Mohammad and Bravo-Marquez, 2017 [65]) to extract both the lexical features and the word embeddings.

Schematic

The schematic of our ensemble is shown in figure 4.1. I have constructed a two layer ensemble. Here are the specifications of the models used:

- SVM (layer 1): `C=0.1, kernel=RBF`
- XGB: `max_depth=5, min_child_weight=150, n_estimators=150, reg_lambda=0.87`
- FFNN (feed forward neural network): Dense (256, sigmoid), Dropout (0.2, sigmoid), Dense (64, sigmoid), Dense (32, sigmoid), Dense (1, relu)
- LSTM-CNN: Conv1D (300, 3, relu), Dropout (0.2), LSTM (150), Dropout (0.2), Dense (32, sigmoid), Dense (1, relu)
- SVM (layer 2): `C=1, kernel=RBF`
- Character level language model (Char): LSTM (150), Dropout (0.2), Dense (64, sigmoid), Dense (1, relu)

Note: XGB stands for the XGBoost implementation of gradient boosted decision trees. SVM was implmented using sklearn (Pedregosa et al., 2011 [78]), while the neural networks were implemented in Keras (Chollet et al., 2015 [25]) with the Tensorflow (Abadi et al., 2016 [2]) backend.

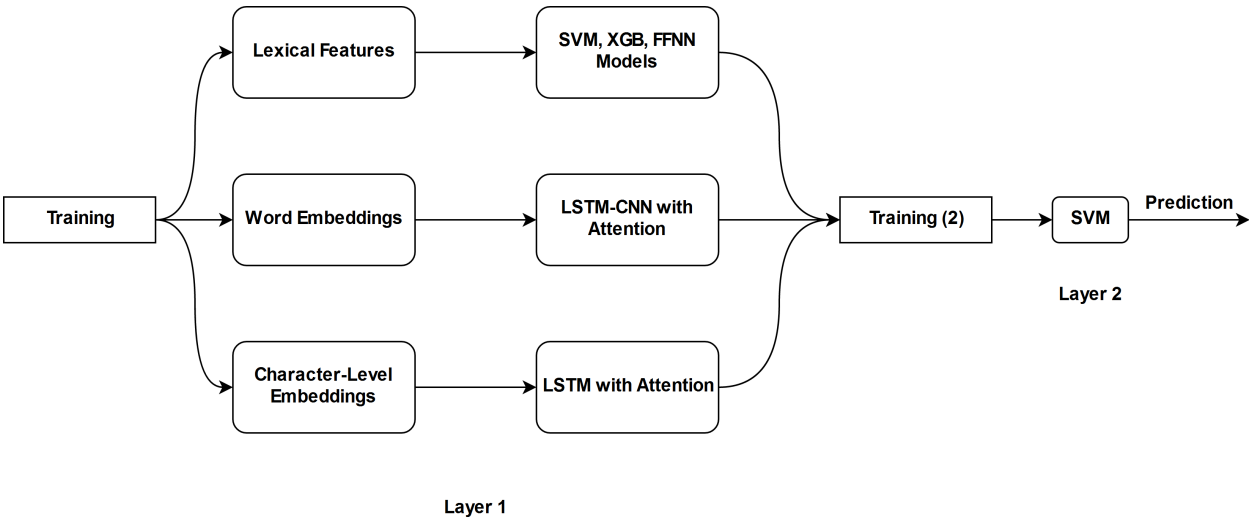


Figure 4.1: The two-layer ensemble for emotion intensity prediction

4.3 RESULTS

I quantify the performance over the SemEval 2018 dataset using the Pearson correlation coefficient between the predictions and the ground truth labels. I first perform an ablation study to understand the relative utility of the various models.

4.3.1 Ablation Study

Given the multiple subsections of data, it is difficult to optimize the architecture and parameters for all emotions for all subtasks. Therefore, I focus on optimizing the architecture and parameters for only the first subtask (emotion intensity prediction) for the emotion *anger*. Given the many models I have developed, it is interesting to see how they perform individually on this subtask. Table 4.2 shows the performance of various feature-model combinations.

I use the **SVM** trained on lexical features as the baseline. We can see that the SVM + XGB + FFNN (referred to as M1) performs better than the SVM alone. LSTM-CNN with attention (referred to as M2) performs similarly to the SVM baseline. However, when combined together, the model M1+M2+Char performs better than each of the individual models on the test set. This means that the different models capture complementary information about the input, and work better in unison, thus demonstrating the efficacy of the idea of ensembling.

Henceforth, I use **M1** to refer to the SVM + XGBoost + Feedforward Neural Network architecture trained on lexical features, **M2** to refer to the LSTM-CNN architecture with

Feature	Model	CV	Dev	Test
L	SVM	0.646	0.616	0.654
	XGB	0.648	0.646	0.634
	FFNN	0.699	0.674	0.664
	SVM+XGB	0.662	0.651	0.663
	SVM+XGB+FFNN [M1]	0.695	0.674	0.673
E	SVM	0.564	0.553	0.555
	LSTM	0.640	0.635	0.633
	LSTM-CNN	0.641	0.639	0.635
	LSTM-CNN (Att) [M2]	0.651	0.642	0.644
L+E	M1+M2	0.733	**0.713**	0.701
	M1+M2+Char	**0.735**	0.711	**0.704**

Table 4.2: An ablation study of various features and models for subtask 1: emotion intensity prediction for the specific case of the emotion *anger*

attention trained on the embedding features and **Char** to refer to the character level LSTM model trained on the individual characters.

4.3.2 Subtasks 1 and 2: Intensity Prediction

Task	Features	Model	Pearson Correlation Coefficient			
			Anger	Fear	Joy	Sadness
1	L	SVM	0.654	0.646	0.649	0.628
	L	M1	0.673	0.668	0.698	0.642
	E	M2	0.644	0.659	0.685	0.644
	L+E	M1+M2+Char	**0.704**	**0.688**	**0.713**	**0.652**
2	L	SVM	0.514	0.449	0.576	0.533
	L	M1	0.549	**0.462**	0.58	0.557
	E	M2	0.544	0.455	0.571	0.542
	L+E	M1+M2+Char	**0.558**	0.461	**0.601**	**0.566**

Table 4.3: Evaluation for subtask 1 (emotion intensity prediction) and subtask 2 (emotion ordinal classification) for all emotions with various features and models

Table 4.3 shows the performance of the models described above to the first two subtasks: emotion intensity prediction and emotion ordinal classification. I have shown the results for all the four emotions. As we can see, here too, the model combination M1+M2+Char combination performs the best for all emotions in subtask 1. The performance of the model is the best for the emotion *joy*, and the worst for the emotion *fear*.

4.3.3 Subtasks 3 and 4: Valence Prediction

Task	Features	Model	Pearson Correlation
3	L	SVM	0.762
	L	M1	0.78
	E	M2	0.764
	L+E	M1+M2+Char	**0.784**
4	L	SVM	0.724
	L	M1	0.733
	E	M2	0.745
	L+E	M1+M2+Char	**0.75**

Table 4.4: Evaluation for subtask 3 (emotion valence regression) and subtask 4 (valence ordinal classification) for various features and models

Coming to subtasks 3 and 4 (valence intensity prediction and valence ordinal classification respectively), table 4.4 shows the performance of the various models on these tasks. Consistent with the results of subtasks 1 and 2, the combined model M1+M2+Char performs the best for both tasks.

Note: In general, we note that the correlation is significantly higher on valence prediction tasks as compared to the emotion intensity tasks. This is likely because the emotion intensity prediction is a fine grained task, requiring the model to observe patterns specific to an emotion. Valence is more of an "aggregated" effect of all the emotions.

4.4 ERROR ANALYSIS

In order to identify areas where the model can improve, it is necessary to study cases where it performs poorly. To do so, I select 5 sentences where the baseline SVM model performs very poorly while predicting anger intensity (based on **absolute error**) and 1 sentence where it performs well. I have restricted the number of sentences to 6 for brevity. In particular, for sentences 1 and 2, the model significantly overestimates the intensity, for sentence 3, the model predicts the intensity accurately. For sentences 4, 5 and 6, the model significantly underestimates the intensity. Table 4.5 shows the sentences considered and the true value of emotion intensity for the emotion anger.

I then compare the absolute error between the true value and model prediction for various models. This comparison is shown in table 4.6. Given that 5 of the 6 sentences are "difficult" for the models, we observe that there is no clear winner over these sentences.

Sentence	Tweet	True Score
1	never had a dull moment with u guys	0.078
2	Fast and furious marathon soon!	0.118
3	They cancelled Chewing Gum. #devastated	0.625
4	Its taking apart my lawn! GET OFF MY LAWN!	0.797
5	I need a beer #irritated	0.806
6	Working with alergies is the most miserable shit in the world #miserable #alergies	0.856

Table 4.5: Test Examples for Error Analysis

However, we observe that for sentences 1 and 2, the model M1 performs relatively well. For sentences 4, 5 and 6, the models *involving* M2 perform relatively well. This suggests that M1 is better at predicting the lower intensities, while M2 is better at the higher intensities. This may explain why though the overall scores for the two models was similar, the ensembled model outperformed the individual models. Another interesting observation is that for sentence 4, the presence of the capital letters is the reason for the high intensity. The model M1+M2+Char is able to identify this well, and contributes to reducing the error significantly as compared to all the other models.

Features	Model	Sentence-wise error					
		1	**2**	**3**	**4**	**5**	**6**
L	SVM	0.310	0.308	**0.004**	-0.377	-0.326	-0.327
L	M1	**0.305**	**0.287**	-0.067	-0.391	-0.286	-0.245
E	M2	0.344	0.366	0.051	-0.265	**-0.241**	**-0.199**
L+E	M1+M2+Char	0.339	0.373	0.071	**-0.213**	-0.242	-0.203

Table 4.6: Absolute error values for various features and models for subtask 1: emotion intensity prediction for emotion *anger*

Chapter 5 Future Work

So far, we have seen the details of the emotion prediction system that I have constructed. In this chapter, I describe some of my future work to improve on this system and to use it to solve other related problems in affective computing.

5.1 IMPROVEMENTS TO THE MODEL

The experimentation reveals that ensembling improves the performance of the system. However, there are a few avenues for improvement based on the error analysis. In particular, I intend to:

- construct emotion-specific models by automating the process of hyperparameter and architecture search
- construct models capturing domain-specific information, for instance, by constructing a separate model for prediction using hashtags
- explore the utility of related architectures, such as the bi-directional LSTM

I also intend to perform a cross-genre study of model performance by evaluating the model over other annotated corpora, such as the news articles (Mihalcea and Liu, 2006 [61]) and children's fairy tales (Alm et al., 2005 [5]).

5.2 ONE-SHOT LEARNING FOR EMOTION DETECTION

Humans possess the ability to learn object categories using only a few training examples at a rapid pace. This problem, termed one shot learning, is common in the computer vision community where the number of object categories can be large (Fei-Fei et al., 2006 [34]). In the domain of computational linguistics, however, I may need to learn a small number of object categories efficiently using few annotated. This problem is especially relevant in emotion detection owing to the cost of annotating data. Therefore, I would like to explore the possibility of using ideas inspired from existing one-shot learning frameworks on the problem of emotion detection.

A popularly used end-to-end architecture for performing one-shot learning is the Siamese network (Koch et al., 2015 [52]). Its inputs are a pair of training samples, with the label being whether the points belong to the same class or not. Two parallel networks generate

a representation of the two inputs, after which the concatenated representations are passed to a classifier that decides whether the pair belongs to the same class or not. This latter classifier learns patterns that distinguish one class from another. An immediate consequence of this architecture is that the effective pairs of labels available to us is approximately the square of the original dataset.

I trained a Siamese network on the task of sentiment classification of IMDb movie reviews. The original dataset consists of 12,500 training and 12,500 test reviews, each containing an approximately equal number of positive and negative reviews. I trained a baseline SVM model with TF-IDF features, and achieved an accuracy of 87.4%. I then constructed a Siamese network with the parallel networks consisting of 2 layer feed-forward networks (300 and 50 neurons respectively) with sigmoid activation. The distinguising network was a fully connected layer with 1 output. Preliminary experiments involved training the network using points sampled from 500 training points in batches of 32. Training over 1000 batches gave a testing accuracy of 83.1%, indicating that the network does perform better than random guessing. I intend to explore the utility of better features (dense embeddings instead of sparse TF-IDF features) and more advanced models (using sequential networks such as LSTMs to learn representations of documents).

5.3 NON-VERBAL EMOTION DETECTION

Written text, especially fiction, contains rich usage of non-verbal expressions. While dialogue is a proven way for a writer to express their thoughts, it tends to be direct and explicit, and may fail to make a reader empathize with the characters. To convey their feelings well, writers often use nonverbal communication. I intend to develop a computational model for non-verbal communication. To do so, I plan to use "The Emotion Thesaurus: A Writer's "Guide To Character Expression"[3]. This thesaurus divides non-verbal emotional indicators into the following three types:

- **Physical signals**: Actions that constitute body language, such as a smile, leaning forward, avoiding eye contact, etc.
- **Internal sensations**: Involuntary responses, such as quickening of the heartbeat, breathlessness, etc.
- **Mental responses** Voluntary but mental reactions to an event, such as jumping to conclusions, fantasizing violence, etc.

For 75 emotions, the authors provide these 3 types of non-verbal indicators along with cues of suppression of each emotion and emotional progression (i.e. what other emotions can the current emotion progress into).

This thesaurus can therefore be used as an effective dataset, as it has several advantages:

- It contains direct descriptions of symptoms of every emotion. I can construct a statistical model centered around the *symptoms* described in this book and enrich it is with world knowledge
- It contains abundant indicators for every emotion
- It has a fine-grained division of emotion. This allows us to view the emotion at various resolutions. I may classify emotion using the classes provided, or I may cluster the emotions together based on progression patterns to get a coarse grained classification.

There are, however, challenges in building such a model. We enumerate some of them below:

- Processing the descriptions has to be performed carefully. Some descriptions (such as *leaning forward*) may be used directly. Others (such as *jealousy towards those interacting with the subject*), less so.
- This thesaurus cannot be used as an exhaustive resource for emotion indicators. I must generalize the model to account for such expressions in language. To do so, I must incorporate a statistical model for world knowledge with the thesaurus. This is a major challenge in this work.

Chapter 6 Conclusion

In this thesis, I have extensively studied the problem of emotion detection in text, and have applied it the domain of Tweets, which have short informal text. I begin by describing the progress in literature in psychology on the subject, by first studying the origins and theories of emotion. I then look at the related work in the domain of computational linguistics to model emotion. I enumerate various models using traditional linguistic features such as word lexicons and the more recent word embedding-based features. I describe both traditional statistical models and non-linear deep neural network based models. I use a combination of these approaches to solve the problem of emotion intensity and valence prediction in SemEval 2018. I construct and evaluate an ensemble model and perform an error analysis over the various models to identify potential sources of improvement to the model.

References

[1] The wikipedia entry for "emoticon", (accessed March 29, 2018). URL `https://en.wikipedia.org/wiki/Emoticon`.

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[3] A. Ackerman and B. Puglisi. *The Emotion Thesaurus: A Writer's Guide to Character Expression*. 2012.

[4] F. H. Allport. A physiological-genetic theory of feeling and emotion. *Psychological Review*, 29(2):132, 1922.

[5] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[6] S. Aman and S. Szpakowicz. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer, 2007.

[7] M. B. Arnold. Physiological differentiation of emotional states. *Psychological Review*, 52(1):35, 1945.

[8] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *LREC*. European Language Resources Association, 2010. ISBN 2-9517408-6-7. URL `http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf`.

[9] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[10] P. Bard. On emotional expression after decortication with some remarks on certain theoretical views: Part i. *Psychological Review*, 41(4):309, 1934.

[11] L. F. Barrett. Emotions are real. *Emotion*, 12(3):413, 2012.

[12] L. F. Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[13] K. Bauman, B. Liu, and A. Tuzhilin. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD*, pages 717–725. ACM, 2017.

[14] G. Bebis and M. Georgiopoulos. Feed-forward neural networks. *IEEE Potentials*, 13 (4):27–31, 1994.

[15] A. C. Boucouvalas. Real time text-to-emotion engine for expressive internet communications. In *Proceedings of International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP-2002)*, 2002.

[16] R. Bougie, R. Pieters, and M. Zeelenberg. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Science*, 31(4):377–393, 2003.

[17] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida, 1999.

[18] M. M. Bradley, M. K. Greenwald, M. C. Petry, and P. J. Lang. Remembering pictures: Pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, 18(2):379, 1992.

[19] F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer. Determining word-emotion associations from tweets by multi-label classification. In *International Conference on Web Intelligence*, pages 536–539. IEEE, 2016.

[20] L. Breiman. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at Berkeley, 1997.

[21] L. Breiman. *Classification and regression trees*. Routledge, 2017.

[22] E. Cambria and A. Hussain. *Sentic computing: Techniques, tools, and applications*, volume 2. Springer Science & Business Media, 2012.

[23] E. Cambria, S. Poria, D. Hazarika, and K. Kwok. Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In *AAAI*, 2018.

[24] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*, pages 785–794. ACM, 2016.

[25] F. Chollet et al. Keras, 2015.

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

[27] C. Darwin. The expression of emotion in animals and man. *London, England: Murray*, 1872.

[28] F. De Waal. *Are we smart enough to know how smart animals are?* WW Norton & Company, 2016.

[29] T. G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

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

[31] P. Ekman. The expression of emotion in men and animals, by charles darwin, 1998.

[32] P. Ekman. What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11(1):31–34, 2016.

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

[34] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[35] K. K. Fitzpatrick, A. Darcy, and M. Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4 (2), 2017.

[36] V. Francisco and P. Gervás. Automated mark up of affective information in english texts. In *International Conference on Text, Speech and Dialogue*, pages 375–382. Springer, 2006.

[37] M. Généreux and R. Evans. Towards a validated model for affective classification of texts. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 55–62. Association for Computational Linguistics, 2006.

[38] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.

[39] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, 2015.

[40] M. Hasan, E. Agu, and E. Rundensteiner. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD Workshop on Health Informatics, New York, USA*, 2014.

[41] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[42] T. K. Ho. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE, 1995.

[43] D. H. Hockenbury and S. E. Hockenbury. *Discovering psychology*. Macmillan, 2010.

[44] L. E. Holzman and W. M. Pottenger. Classification of emotions in internet chat: An application of machine learning using speech phonemes. *Retrieved November*, 27(2011): 50, 2003.

[45] J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[46] C. E. Izard. The face of emotion. 1971.

[47] A. Kalra and K. Karahalios. Texttone: expressing emotion through text. In *IFIP Conference on Human-Computer Interaction*, pages 966–969. Springer, 2005.

[48] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke. *Words with attitude*. Citeseer, 2001.

[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3149–3157, 2017.

[50] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[51] S. Kiritchenko, X. Zhu, and S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[52] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[53] Z. Kövecses. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press, 2003.

[54] C. G. Lange and W. James. *The emotions*, volume 1. Williams & Wilkins, 1922.

[55] R. S. Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819, 1991.

[56] D. J. Litman and K. Forbes-Riley. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 351. Association for Computational Linguistics, 2004.

[57] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM, 2003.

[58] G. D. Marshall and P. G. Zimbardo. Affective consequences of inadequately explained physiological arousal. 1979.

[59] A. Mehrabian. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. 1980.

[60] M. Michailidis. Stacknet meta modelling framework. `https://github.com/kaz-Anova/StackNet`, 2017. Accessed: 2018-02-26.

[61] R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144, 2006.

[62] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[63] Y. Miyamoto, Y. Uchida, and P. C. Ellsworth. Culture and mixed emotions: Co-occurrence of positive and negative emotions in japan and the united states. *Emotion*, 10(3):404, 2010.

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

[65] S. Mohammad and F. Bravo-Marquez. Wassa shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 34–49, 01 2017.

[66] S. M. Mohammad and F. Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark, 2017.

[67] S. M. Mohammad and F. Bravo-Marquez. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696*, 2017.

[68] S. M. Mohammad and F. Bravo-Marquez. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*, 2017.

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

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

[71] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA*, 2018.

[72] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Sentiful: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

[73] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[74] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[75] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[76] W. G. Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[77] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[79] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[80] S. Petrović, M. Osborne, and V. Lavrenko. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.

[81] R. Plutchik. The multifactor-analytic theory of emotion. *the Journal of Psychology*, 50(1):153–171, 1960.

[82] R. Plutchik. The emotions: Facts. *Theories and a New Model, New York*, 1962.

[83] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[84] D. Preotiuc-Pietro, H. A. Schwartz, G. Park, J. C. Eichstaedt, M. Kern, L. Ungar, and E. P. Shulman. Modelling valence and arousal in facebook posts. In *Proceedings of NAACL-HLT*, pages 9–15, 2016.

[85] N. A. Remington, L. R. Fabrigar, and P. S. Visser. Reexamining the circumplex model of affect. *Journal of personality and social psychology*, 79(2):286, 2000.

[86] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

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

[88] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[89] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[90] S. Schachter and J. Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379, 1962.

[91] H. Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81, 1954.

[92] C. A. Smith and L. D. Kirby. Putting appraisal in context: Toward a relational model of appraisal and emotion. *Cognition and Emotion*, 23(7):1352–1372, 2009.

[93] J. Staiano and M. Guerini. Depechemood: A lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

[94] P. J. Stone, R. F. Bales, J. Z. Namenwirth, and D. M. Ogilvie. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498, 1962.

[95] C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.

[96] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[97] C. Strapparava, A. Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086. Citeseer, 2004.

[98] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy systems*, 9(4):483–496, 2001.

[99] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[100] M. Tan, C. d. Santos, B. Xiang, and B. Zhou. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*, 2015.

[101] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[102] J. Velsquez. Modeling emotions and other motivations in synthetic agents. In *AAAI*, volume 97, pages 10–15, 1997.

[103] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015.

[104] S. Vosoughi, P. Vijayaraghavan, and D. Roy. Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1041–1044. ACM, 2016.

[105] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.

[106] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

[107] D. Watson and A. Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219, 1985.

[108] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

[109] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*, pages 347–354. Association for Computational Linguistics, 2005.

[110] W. Wundt. Outlines of psychology, trans. by ch judd (engelmann). *Original work published*, 1896.

[111] Y. Zhang and B. Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

[112] X. Zhe and A. Boucouvalas. Text-to-emotion engine for real time internet communication. In *Proceedings of International Symposium on Communication Systems, Networks and DSPs*, pages 164–168. Citeseer, 2002.