Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text

Saif M. Mohammad SAIF.MOHAMMAD@NRC-CNRC.GC.CA
National Research Council Canada
1200 Montreal Rd., Ottawa, ON, Canada

1. Introduction

The term *sentiment analysis* can be used to refer to many different, but related, problems. Most commonly, it is used to refer to the task of automatically determining the valence or polarity of a piece of text, whether it is positive, negative, or neutral. However, more generally, it refers to determining one's attitude towards a particular target or topic. Here, attitude can mean an evaluative judgment, such as positive or negative, or an emotional or affectual attitude such as frustration, joy, anger, sadness, excitement, and so on. Note that some authors consider *feelings* to be the general category that includes attitude, emotions, moods, and other affectual states. In this chapter, we use 'sentiment analysis' to refer to the task of automatically determining feelings from text, in other words, automatically determining valence, emotions, and other affectual states from text.

Osgood, Suci, and Tannenbaum (1957) showed that the three most prominent dimensions of meaning are evaluation (*good–bad*), potency (*strong–weak*), and activity (*active–passive*). Evaluativeness is roughly the same dimension as valence (*positive–negative*). Russell (1980) developed a circumplex model of affect characterized by two primary dimensions: valence and arousal (degree of reactivity to stimulus). Thus, it is not surprising that large amounts of work in sentiment analysis are focused on determining valence. (See survey articles by Pang and Lee (2008), Liu and Zhang (2012), and Liu (2015).) However, there is some work on automatically detecting arousal (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010; Kiritchenko, Zhu, & Mohammad, 2014b; Mohammad, Kiritchenko, & Zhu, 2013a) and growing interest in detecting emotions such as anger, frustration, sadness, and optimism in text (Mohammad, 2012; Bellegarda, 2010; Tokuhisa, Inui, & Matsumoto, 2008; Strapparava & Mihalcea, 2007; John, Boucouvalas, & Xu, 2006; Mihalcea & Liu, 2006; Genereux & Evans, 2006; Ma, Prendinger, & Ishizuka, 2005; Holzman & Pottenger, 2003; Boucouvalas, 2002; Zhe & Boucouvalas, 2002). Further, massive amounts of data emanating from social media have led to significant interest in analyzing blog posts, tweets, instant messages, customer reviews, and Facebook posts for both valence (Kiritchenko et al., 2014b; Kiritchenko, Zhu, Cherry, & Mohammad, 2014a; Mohammad et al., 2013a; Aisopos, Papadakis, Tserpes, & Varvarigou, 2012; Bakliwal, Arora, Madhappan, Kapre, Singh, & Varma, 2012; Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Thelwall, Buckley, & Paltoglou, 2011; Brody & Diakopoulos, 2011; Pak & Paroubek, 2010) and emotions (Hasan, Rundensteiner, & Agu, 2014; Mohammad & Kiritchenko, 2014; Mohammad, Zhu, Kiritchenko, & Martin, 2014; Choudhury, Counts, & Gamon, 2012; Mohammad, 2012a; Wang, Chen, Thirunarayan, & Sheth, 2012; Tumasjan, Sprenger, Sandner, & Welpe, 2010b; Kim, Gilbert, Edwards, &

Graeff, 2009; Bollen, Pepe, & Mao, 2009; Aman & Szpakowicz, 2007). Ortony, Clore, and Collins (1988) argue that all emotions are valenced, that is, emotions are either positive or negative, but never neutral (Ortony et al., 1988). While instantiations of some emotions tend to be associated with exactly one valence (for example, joy is always associated with positive valence), instantiations of other emotions may be associated with differing valence (for example, some instances of surprise are associated with positive valence, while some others are associated with negative valence). Thus, methods for emotion classification often benefit from using valence features. The vast majority of these valence and emotion classification approaches employ statistical machine learning techniques, although some rule-based approaches, such as Neviarouskaya, Prendinger, and Ishizuka (2009), also persist.

Automatic detection and analysis of affectual categories in text has wide-ranging applications. Below we list some key directions of ongoing work:

- Public Health: Automatic methods for detecting emotions are useful in detecting depression (Pennebaker, Mehl, & Niederhoffer, 2003; Rude, Gortner, & Pennebaker, 2004; Cherry, Mohammad, & De Bruijn, 2012), identifying cases of cyber-bullying (Chen, Zhou, Zhu, & Xu, 2012; Dadvar, Trieschnigg, Ordelman, & de Jong, 2013), predicting health attributes at community level (Johnsen, Vambheim, Wynn, & Wangberg, 2014; Eichstaedt, Schwartz, Kern, Park, Labarthe, Merchant, Jha, Agrawal, Dziurzynski, Sap, et al., 2015), and tracking well-being (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, Agrawal, Park, et al., 2013; Paul & Dredze, 2011). There is also interest in developing robotic assistants and physio-therapists for the elderly, the disabled, and the sick—robots that are sensitive to the emotional state of the patient.

- Politics: There is tremendous interest in tracking public sentiment, especially in social media, towards politicians, electoral issues, as well as national and international events. Some studies have shown that the more partisan electorate tend to tweet more, as do members from minority groups (Lassen & Brown, 2011). There is work on identifying contentious issues (Maynard & Funk, 2011) and on detecting voter polarization (Conover, Ratkiewicz, Francisco, Gonc, Flammini, & Menczer, 2011a). Tweet streams have been shown to help identify current public opinion towards the candidates in an election (nowcasting) (Golbeck & Hansen, 2011; Conover, Goncalves, Ratkiewicz, Flammini, & Menczer, 2011b; Mohammad et al., 2014). Some research has also shown the predictive power of analyzing electoral tweets to determine the number of votes a candidate will get (forecasting) (Tumasjan, Sprenger, Sandner, & Welpe, 2010a; Bermingham & Smeaton, 2011; Lampos, Preotiuc-Pietro, & Cohn, 2013). However, other research expresses skepticism at the extent to which forecasting is possible (Avello, 2012).

- Brand management, customer relationship management, and Stock market: Sentiment analysis of blogs, tweets, and Facebook posts is already widely used to shape brand image, track customer response, and in developing automatic dialogue systems for handling customer queries and complaints (Ren & Quan, 2012; Yen, Lin, & Lin, 2014; Yu, Wu, Chang, & Chu, 2013; Gupta, Gilbert, & Fabbrizio, 2013; Fang, Chen, Wang, & Wu, 2011; Bock, Gluge, Wendemuth, Limbrecht, Walter, Hrabal, & Traue, 2012).

- Education: Automatic tutoring and student evaluation systems detect emotions in responses to determine correctness of responses and also to determine emotional state of the participant (Li, Li, Jiang, & Zhang, 2014; Suero Montero & Suhonen, 2014). It has been shown that learning improves when the student is in a happy and calm state as opposed to anxious or frustrated (Dogan, 2012).

- Tracking The Flow of Emotions in Social Media: Besides work in brand management and public health, as discussed already, some recent work attempts to better understand how emotional information *spreads* in a social network, for instance to improve disaster management (Kramer, 2012; Vo & Collier, 2013).

- Detecting Personality Traits: Systematic patterns in how people express emotions is a key indicator of personality traits such as extroversion and narcissism. Thus many automatic systems that determine personality traits from written text rely on automatic detection of emotions (Grijalva, Newman, Tay, Donnellan, Harms, Robins, & Yan, 2014; Minamikawa & Yokoyama, 2011; Schwartz et al., 2013; Malti & Krettenauer, 2013; Mohammad & Kiritchenko, 2013).

- Understanding Gender Differences: Men and woman use different language socially, at work, and even in computer-mediated communication. Several studies have analyzed the differences in emotions in language used by men and women in these contexts (Grijalva et al., 2014; Montero, Munezero, & Kakkonen, 2014; Mohammad & Yang, 2011a).

- Literary Analysis: There is growing interest in using automatic natural language processing techniques to analyze large collections of literary texts. Specifically with respect to emotions, there is work on tracking the flow of emotions in novels, plays, and movie scripts, detecting patterns of sentiment common to large collections of texts, and tracking emotions of plot characters (Hartner, 2013; Kleres, 2011; Mohammad, 2011, 2012b). There is also work in generating music that captures the emotions in text (Davis & Mohammad, 2014).

- Visualizing Emotions: A number of applications listed above benefit from good visualizations of emotions in text(s). Particularly useful is the feature of interactivity. If users are able to select particular aspects such as an entity, emotion, or time-frame of interest, and the system responds to show information relevant to the selection in more detail, then the visualization enables improved user-driven exploration of the data. Good visualizations also help users gain new insights and can be a tool for generating new ideas. See Quan and Ren (2014), Mohammad (2012b), Liu, Selker, and Lieberman (2003b), Gobron, Ahn, Paltoglou, Thelwall, and Thalmann (2010) for work on visualization of emotions in text.

As automatic methods to detect various affect categories become more accurate, their use in natural language applications will likely become even more ubiquitous.

This chapter presents a comprehensive overview of work on automatically detecting sentiment in text.[1] We begin in Section 2 by discussing various challenges to sentiment

1. See surveys by El Ayadi, Kamel, and Karray (2011) and Anagnostopoulos, Iliou, and Giannoukos (2015) for an overview of emotion detection in speech. See Picard (2000) and Alm (2008) for a broader intro-

analysis. In Section 3, we describe the diverse landscape of sentiment analysis problems, including: detecting sentiment of the writer, reader, and other relevant entities; detecting sentiment from words, sentences, and documents; detecting stance towards events and entities which may or may not be explicitly mentioned in the text; detecting sentiment towards aspects of products; and detecting semantic roles of feelings.

Many of the machine learning approaches for automatic detection of sentiment are supervised, that is, systems first learn a model from a set of example instances labeled with the correct sentiment. (This set of example instances is called the *training set*.) Then the model is able to predict the sentiment of new, previously unseen, instances. To determine the expected prediction accuracy, the model is often evaluated on a held-out portion of the labeled data called the *test set*. (There is no overlap between the instances in the training and test sets.) In Section 4, we discuss work on creating labeled data (training and test sets) for valence and emotion. We also summarize automatic methods to detect valence and emotion in text. Many of these approaches rely on lists of words associated with affect categories. We describe approaches to create large lexicons of term–affect associations in Section 5. Some classes of terms, such as negation words and degree adverbs, are not directly associated with sentiment, but they impact the sentiment of other terms in their vicinity. Section 6 describes work on modeling the impact of modifiers such as negation and degree adverbs on sentiment.

Language is rife with creativity in the form of metaphors, analogies, expressions of sarcasm, statements of irony, and so on. Such texts are collectively referred to as *figurative language* in the natural language processing community, and they are especially challenging for automatic text analysis systems. Section 7 discusses some preliminary sentiment analysis work focused on figurative language.

Since much of the research and resource development in sentiment analysis has been on English texts, sentiment analysis systems in other languages tend to be less accurate. This has ushered work in leveraging the resources in English for sentiment analysis in the resource poor languages. We discuss this work in Section 8.

Automatic text analysis systems are often evaluated on different datasets and with different settings. Thus the results reported in different articles are often not directly comparable, making it hard to assess which approach is better in practice. Further, machine learning algorithms can often "overfit" on the training data, that is, a system may be predict accurately on test data that is very similar to the training data, but poorly on test data that is from a different domain or from a different time span. To address this, shared task competitions are organized to evaluate various algorithms on a common evaluation framework, with new, never before seen, datasets. Many of the sentiment analysis related shared tasks were organized under the aegis of *SemEval (Semantic Evaluation)* – a two-day workshop, normally held in conjunction with a Natural Language Processing (NLP) conference.[2] Throughout the chapter we highlight some of the best approaches in sentiment analysis, including approaches that were most successful in these shared task competitions organized by the NLP community. Finally, in Section 9, we present future directions.

duction of giving machines the ability to detect sentiment and emotions in various modalities such as text, speech, and vision.

2. http://aclweb.org/aclwiki/index.php?title=SemEval_Portal

2. Challenges in Sentiment Analysis

There are several challenges to automatically detecting sentiment in text:
Complexity and Subtlety of Language Use:

- The emotional import of a sentence or utterance is not simply the sum of emotional associations of its component words. Further, Emotions are often not explicitly stated. For example:

  Another Monday, and another week working my tail off.

  conveys a sense of frustration without the speaker explicitly saying so. Note that the sentence does not include any overtly negative words.
  Section 4 summarizes various machine learning approaches for classifying sentences and tweets into one of the affect categories.

- Certain terms such as negations and modals impact sentiment of the sentence, without themselves having strong sentiment associations. For example, *may be good*, *was good*, and *was not good* should be interpreted differently by sentiment analysis systems.
  Section 6 discusses approaches that explicitly handle sentiment modifiers such as negations, degree adverbs, and modals.

- Words when used in different contexts (and different senses) can convey different emotions. For example, the word *hug* in the *embrace* sense, as in:

  Mary hugged her daughter before going to work.

  is associated with joy and affection, but *hug* in the *stay close to sense*, as in:

  The pipeline hugged the state border.

  is rather unemotional. Word sense disambiguation remains a difficult challenge in natural language processing (Kilgarriff, 1997; Navigli, 2009).
  In Section 5, we discuss approaches to create term–sentiment association lexicons, including some that have separate entries for each sense of a word.

- Utterances may convey more than one emotion (and to varying degrees). They may convey contrastive evaluations of multiple target entities.

- Utterances may refer to emotional events without implicitly or explicitly expressing the feelings of the speaker.

Use of Creative and Non-Standard Language:

- Automatic natural language systems find it difficult to interpret creative uses of language such as sarcasm, irony, humour, and metaphor. However, these phenomenon are common in language use.
  Section 7 summarizes some preliminary work in this direction.

- Social media texts are rife with terms not seen in dictionaries such as misspellings (*parlament*), creatively-spelled words (*happeee*), hashtagged words (*#loveumom*), emoticons, abbreviations (*lmao*), etc. Many of these terms convey emotions.
  Section 5.2 describes work on automatically generating term–sentiment association lexicons from social media data—methods that capture sentiment associations of not

just regular English terms, but also social media specific terms.

Lack of Para-Linguistic Information:

- Often we communicate affect through tone, pitch, and emphasis. However, written text usually does not come with annotations of stress and intonation. This is compensated to some degree by the use of explicit emphasis markers (for example, Mary used *Jack's* computer) and explicit sentiment markers such as emoticons and emoji.

- We also communicate emotions through facial expressions. In fact there is a lot of work linking different facial expressions to different emotional states (Ekman, 1992; Ekman & Friesen, 2003). (Also, see Chapter 11 by Hwang and Matsumoto in this book.) Once again, this information is not present in written text.

Lack of Large Amounts of Labeled Data:

- Most machine learning algorithms for sentiment analysis require significant amounts of training data (example sentences marked with the associated emotions). However, there are numerous affect categories including hundreds of emotions that humans can perceive and express. Thus, much of the work in the community has been restricted to a handful of emotions and valence categories.
  Section 4.3 summarizes various efforts to create datasets that have sentences labeled with emotions.

Subjective and Cross-Cultural Differences:

- Detecting emotions in text can be difficult even for humans. Studies have shown that the amount of agreement between annotators is significantly lower in assigning valence or emotions to instances, as compared to tasks such as identifying part of speech and detecting named entities.

- There can be significant differences in emotions associated with events and behaviors across different cultures. For example, *dating* and *alcohol* may be perceived as significantly more negative in some parts of the world than in others.

- Manual annotations can be significantly influenced by clarity of directions, difficulty of task, training of the respondents, and even the annotation scheme (multiple choice questions, free text, Likert scales, etc.).
  Sections 4 and 5 describe various manually annotated datasets where affect labels are provided for sentences and words, respectively. They were created either by hand-chosen expert annotators, known associates and grad students, or by crowdsourcing on the Internet to hundreds or thousands of unknown respondents. Section 5.1.1 describes an annotation scheme called *maximum difference scaling (MaxDiff)* or *best-worst scaling* (Louviere, 1991) that has led to more high-quality and consistent sentiment annotations.

In the sections ahead we describe approaches that, to some extent, address these issues. Nonetheless, significant challenges still remain. Additionally, sentiment analysis involves a diverse landscape of tasks, each of which can be operationalized in multiple ways.

3. Sentiment Analysis Tasks

3.1 Detecting Sentiment of the Writer, Reader, and other Entities

Sentiment can be associated with any of the following: 1. the speaker (or writer), 2. the listener (or reader), or 3. one or more entities mentioned in the utterance. Most research in sentiment analysis has focused on detecting the sentiment of the speaker, and this is often done by analyzing only the utterance. However, there are several instances where it is unclear whether the sentiment in the utterance is the same as the sentiment of the speaker. For example, consider:

Sarah: *The war in Syria has created a refugee crisis.*

The sentence describes a negative event (millions of people being displaced), but it is unclear whether to conclude that Sarah (the speaker) is personally saddened by the event. It is possible the Sarah is a news reader and merely communicating information about the war. Developers of sentiment systems have to decide before hand whether they wish to assign a negative sentiment or neutral sentiment to the speaker in such cases. More generally, they have to decide whether the speaker's sentiment will be chosen to be neutral in absence of clear signifiers of the speaker's own sentiment, or whether the speaker's sentiment will be chosen to be the same as the sentiment of events and topics mentioned in the utterance.

On the other hand, people can react differently to the same utterance, for example, people on opposite sides of a debate or rival sports fans. Thus modeling listener sentiment requires modeling listener profiles. This is an area of research not explored much by the community. Similarly, there is little work on modeling sentiment of entities mentioned in the text, for example, given:

Drew: *Jamie could not stop gushing about the new Game of Thrones episode.*

It will be useful to develop automatic systems that can deduce that Jamie (not Drew) liked the new episode of *Game of Thrones* (a TV show).

3.2 Detecting Sentiment from Different Textual Chunks

Sentiment can be determined at various levels: from sentiment associations of words and phrases; to sentiments of sentences, SMS messages, chat messages, and tweets; to sentiments in product reviews, blog posts, and whole documents.

Words: Some words signify valence as part of their core meaning, for example, *good, bad, terrible, excellent, nice,* and so on. Some other words do not signify valence as part of their meaning, but have strong associations with positive or negative valence. For example, *party* and *raise* are associated with positive valence, whereas *slave* and *death* are associated with negative valence.[3] Words that are not strongly associated with positive or negative

3. Note that words that signify valence, are also associated with that valence, but words that are associated with a valence, do not always signify that valence as part of their meaning.

valence are considered neutral. (The exact boundaries between neutral and positive valence, and between neutral and negative valence, are somewhat fuzzy. However, for a number of terms, there is high inter-rater agreement on whether they are positive, neutral, or negative.) Similarly, some words express emotions as part of their meaning (and thus are also associated with the emotion), and some words are just associated with emotions. For example, *anger* and *rage* denote anger (and are associated with anger), whereas *negligence, fight,* and *betrayal* do not denote anger, but they are associated with anger.

Sentiment associations of words and phrases are commonly captured in valence and emotion association lexicons. A valence (or polarity) association lexicon may have entries such as these shown below (text in parenthesis is not part of the entry, but our description of what the entry indicates):

delighted – positive (*delighted* is usually associated with positive valence)
death – negative (*death* is usually associated with negative valence)
shout – negative (*shout* is usually associated with negative valence)
furniture – neutral (*furniture* is **not** strongly associated with positive or negative valence)

An affect association lexicon has entries for a pre-decided set of emotions (different lexicons may choose to focus on different sets of emotions). Below are examples of some affect association entries:

delighted – joy (*delighted* is usually associated with the emotion of joy)
death – sadness (*death* is usually associated with the emotion of sadness)
shout – anger (*shout* is usually associated with the emotion of anger)
furniture – none (*furniture* is **not** strongly associated with any of the pre-decided set of emotions)

A word may be associated with more than one emotion, in which case, it will have more than one entry in the affect lexicon.

Sentiment association lexicons can be created either by manual annotation or through automatic means. Manually created lexicons tend to be in the order of a few thousand entries, but automatically generated lexicons can capture valence and emotion associations for hundreds of thousands unigrams (single word strings) and even for larger expressions such as bigrams (two-word sequences) and trigrams (three-word sequences). Automatically generated lexicons often also include a real-valued score indicating the strength of association between the word and the affect category. This score is the prior estimate of the sentiment association, calculated from previously seen usages of the term. While sentiment lexicons are often useful in sentence-level sentiment analysis, the same terms may convey different sentiments in different contexts. The top systems (Mohammad et al., 2013a; Kiritchenko et al., 2014a; Zhu, Kiritchenko, & Mohammad, 2014b; Tang, Wei, Qin, Liu, & Zhou, 2014a) in recent sentiment-related shared tasks, SemEval-2013 and 2014 Sentiment Analysis in Twitter, used large sentiment lexicons (Wilson, Kozareva, Nakov, Rosenthal, Stoyanov, & Ritter, 2013; Rosenthal, Nakov, Ritter, & Stoyanov, 2014).[4] The tasks also had separate

4. https://www.cs.york.ac.uk/semeval-2013/
http://alt.qcri.org/semeval2014/

sub-tasks aimed at identifying sentiment of terms in context. We discuss manually and automatically created valence and emotion association lexicons in more detail in Section 5.

Sentences: Sentence-level valence classification systems assign labels such as positive, negative, or neutral to whole sentences. It should be noted that the valence of a sentence is not simply the sum of the polarities of its constituent words. Automatic systems learn a model from labeled training data (instances that are already marked as positive, negative, or neutral) using a large number of features such as word and character ngrams, valence association lexicons, negation lists, word clusters, and, more recently, features from low-dimensional vector representations of words. We discuss these approaches in Section 4.2.

Emotion classification systems assign labels such as joy, sadness, anger, and fear to sentences. They too use feature sets similar to the valence classification systems. In contrast to valence classification (for which there have been many), there has been only one shared task competition on detecting emotions—the 2007 SemEval competition *Affective News* (Strapparava & Mihalcea, 2007), where participants had to determine the emotions in newspaper headlines.[5] This was framed as an unsupervised task. The competition drew a small number of participants, most of which failed to surpass the most-frequent class baseline (the accuracy obtained by always guessing the most frequent class in the dataset). However, methods of garnering supplemental training data by clever methods such as distant supervision have led to more progress in the area. We discuss some of these approaches in Section 4.3.

Documents: Sentiment analysis of documents is often broken down into the sentiment analysis of the component sentences. Thus we do not discuss this topic in much detail here. However, there is interesting work on using sentiment analysis to generate text summaries (Ku, Liang, & Chen, 2006; Liu, Cao, Lin, Huang, & Zhou, 2007; Somprasertsri & Lalitrojwong, 2010; Stoyanov & Cardie, 2006; Lloret, Balahur, Palomar, & Montoyo, 2009) and on analyzing patterns of sentiment in social networks in novels and fairy tales (Nalisnick & Baird, 2013b, 2013a; Mohammad & Yang, 2011b; Davis & Mohammad, 2014).

3.3 Detecting Sentiment Towards a Target

3.3.1 Detecting Sentiment Towards Aspects of an Entity

A review of a product or service can express sentiment towards various aspects. For example, a restaurant review can gush positively about the food, but express anger towards the quality of service. There is now a growing amount of work in detecting aspects of products and also sentiment towards these aspects (Popescu & Etzioni, 2005; Su, Xiang, Wang, Sun, & Yu, 2006; Xu, Huang, & Wang, 2013; Qadir, 2009; Zhang, Liu, Lim, & O'Brien-Strain, 2010; Kessler & Nicolov, 2009). In 2014, a shared task was organized for detecting aspect sentiment in restaurant and laptop reviews (Pontiki, Galanis, Pavlopoulos, Papageorgiou, Androutsopoulos, & Manandhar, 2014). The best performing systems had a strong sentence-level sentiment analysis system to which they added localization features so that more weight was given to sentiment features close to the mention of the aspect. This task was repeated in 2015. It will be useful to develop aspect-based sentiment systems for other domains such

5. http://web.eecs.umich.edu/∼mihalcea/downloads.html#affective

as blogs and news articles as well. (See proceedings of SemEval-2014 and 2015 for details about participating aspect sentiment systems.[6])

3.3.2 Detecting Stance

Stance detection is the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target. For example, given the following target and text pair:

Target: *women have the right to abortion*
Text: *A foetus has rights too!*

Humans can deduce from the text that the speaker is against the proposition. However, this is a challenging task for computers. To successfully detect stance, automatic systems often have to identify relevant bits of information that may not be present in the focus text. For example, that if one is actively supporting foetus rights, then he or she is likely against the right to abortion. Automatic systems can obtain such information from large amounts of text about the target.

Stance detection is related to sentiment analysis, but the two have significant differences. In sentiment analysis, systems determine whether a piece of text is positive, negative, or neutral. However, in stance detection, systems are to determine favorability towards a given target – and the target may not be explicitly mentioned in the text. For example, consider the target–text pair below:

Target: *Barack Obama*
Text: *Romney will be a terrible president.*

The tweet was posted during the 2012 US presidential campaign between Barack Obama and Mitt Romney. Note that the text is negative in sentiment (and negative towards Mitt Romney), but the tweeter is likely to be favorable towards the given target (Barack Obama). Also note that one can be against Romney but not in favor of Obama, but in stance detection, the goal is to determine which is more probable: that the author is in favour of, against, or neutral towards the target.

Automatically detecting stance has widespread applications in information retrieval, text summarization, and textual entailment. In fact, one can argue that stance detection can bring complementary information to sentiment analysis, because we often care about the authors evaluative outlook towards *specific targets* and propositions rather than simply about whether the speaker was angry or happy.

Over the last decade, there has been active research in modeling stance. However, most works focus on congressional debates (Thomas, Pang, & Lee, 2006) or debates in online forums (Somasundaran & Wiebe, 2009; Murakami & Raymond, 2010; Anand, Walker, Abbott, Tree, Bowmani, & Minor, 2011; Walker, Anand, Abbott, & Grant, 2012; Hasan & Ng, 2013; Sridhar, Getoor, & Walker, 2014). New research in domains such as social media texts, and approaches that combine traditional sentiment analysis with relation extraction can make a significant impact in improving the state-of-the-art in automatic stance detection.

6. http://alt.qcri.org/semeval2014/
http://alt.qcri.org/semeval2015/

3.4 Detecting Semantic Roles of Emotion

The Theory of Frame Semantics argues that the meanings of most words can be understood in terms of a set of related entities and their relations (Fillmore, 1976, 1982). For example, the concept of education usually involves a student, a teacher, a course, an institution, duration of study, and so on. The set of related entities is called a *semantic frame* and the individual entities, defined in terms of the role they play with respect to the target concept, are called the *semantic roles*. *FrameNet* (Baker, Fillmore, & Lowe, 1998) is a lexical database of English that records such semantic frames.[7] Table 1 shows the FrameNet frame for emotions. Observe that the frame depicts various roles such as who is experiencing the emotion (the *experiencer*), the person or event that evokes the emotion, and so on. Information retrieval, text summarization, and textual entailment benefit from determining not just the emotional state but also from determining these semantic roles of emotion.

Mohammad, Zhu, Kiritchenko, and Martin (2015) created a corpus of tweets from the run up to the 2012 US presidential elections, with annotations for valence, emotion, stimulus, and experiencer. The tweets were also annotated for intent (to criticize, to support, to ridicule, etc.) and style (simple statement, sarcasm, hyperbole, etc.). The dataset is made available for download.[8] They also show that emotion detection alone can fail to distinguish between several different types of intent. For example, the same emotion of disgust can be associated with the intents of 'to criticize', 'to vent', and 'to ridicule'. They also developed systems that automatically classify electoral tweets as per their emotion and purpose, using various features that have traditionally been used in tweet classification, such as word and character ngrams, word clusters, valence association lexicons, and emotion association lexicons. Ghazi, Inkpen, and Szpakowicz (2015) compiled FrameNet sentences that were tagged with the stimulus of certain emotions. They also developed a statistical model to detect spans of text referring to the emotion stimulus.

4. Detecting Subjectivity, Valence, and Emotions in Sentences and Tweets

Sentiment analysis systems have been applied to many different kinds of texts including customer reviews (Pang & Lee, 2008; Liu & Zhang, 2012; Liu, 2015), news paper headlines (Bellegarda, 2010), novels (Boucouvalas, 2002; John et al., 2006; Francisco & Gervás, 2006; Mohammad & Yang, 2011b), emails (Liu, Lieberman, & Selker, 2003a; Mohammad & Yang, 2011b), blogs (Neviarouskaya et al., 2009; Genereux & Evans, 2006; Mihalcea & Liu, 2006), and tweets (Pak & Paroubek, 2010; Agarwal et al., 2011; Thelwall et al., 2011; Brody & Diakopoulos, 2011; Aisopos et al., 2012; Bakliwal et al., 2012; Mohammad, 2012a). Often the analysis of documents and blog posts is broken down into determining the sentiment within each component sentence. In this section we discuss approaches for such sentence-level sentiment analysis. Even though tweets may include more than one sentence, they are limited to 140 characters, and most are composed of just one sentence. Thus we include here work on tweets as well.

7. https://framenet.icsi.berkeley.edu/fndrupal/home
8. Political Tweets Dataset: www.purl.org/net/PoliticalTweets

Role	Description
Core:	
Experiencer	the person that experiences or feels the emotion
State	the abstract noun that describes the experience
Stimulus	the person or event that evokes the emotional response
Topic	the general area in which the emotion occurs
Non-Core:	
Circumstances	the condition in which Stimulus evokes response
Degree	The extent to which the Experiencer's emotion deviates from the norm for the emotion
Empathy_target	The Empathy_target is the individual or individuals with which the Experiencer identifies emotionally
Manner	Any description of the way in which the Experiencer experiences the Stimulus which is not covered by more specific frame elements
Reason	the explanation for why the Stimulus evokes an emotional response

Table 1: The FrameNet frame for emotions.

4.1 Detecting Subjectivity

One of the earliest problems tackled in sentiment analysis is that of detecting subjective language (Wiebe, Wilson, Bruce, Bell, & Martin, 2004; Wiebe & Riloff, 2005). For example, sentences can be classified as subjective (having opinions and attitude) or objective (containing facts). This has applications in question answering, information retrieval, paraphrasing, and other natural language applications where it is useful to separate factual statements from speculative or affectual ones. For example, if the target query is "*what did the users think of iPhone 5's screen?*", then the question answering system (or information retrieval system) should be able to distinguish between sentences such as *"the iPhone has a beautiful touch screen"* and sentences such as *"iPhone 5 has 326 pixels per inch"*. Sentences like the former which express opinion about the screen should be extracted. On the other hand, if the user query is "*what is iPhone 5's screen resolution?*", then sentences such as the latter (referring to 326 pixels per inch) are more relevant. (See Wiebe and Riloff (2011) for work on using subjectivity detection in tandem with techniques for information extraction.) It should be noted, however, that if a sentence is objective, then it does not imply that the sentence is necessarily true. It only implies that the sentence does not exhibit the speaker's private state (attitude, evaluations, and emotions). Similarly, if a sentence is subjective, that does not imply that it lacks truth.

A number of techniques have been proposed to detect subjectivity using patterns of word usage, identifying certain kinds of adjectives, detecting emotional terms, and occurrences of certain discourse connectives (Hatzivassiloglou & Wiebe, 2000; Riloff & Wiebe, 2003; Wiebe et al., 2004; Pit, 2006; Su & Markert, 2008; Das & Bandyopadhyay, 2009; Lin, He, & Everson, 2011; Wang & Fu, 2010). Opinion Finder is one of the most popular freely available subjectivity systems (Wilson, Hoffmann, Somasundaran, Kessler, Wiebe, Choi, Cardie, Riloff, & Patwardhan, 2005).[9]

9. http://mpqa.cs.pitt.edu/opinionfinder/

4.2 Detecting Valence

There is tremendous interest in accurately determining valence in sentences and tweets. Surveys by Pang and Lee (2008), Liu and Zhang (2012), and Martínez-Cámara, Martín-Valdivia, Ureñalópez, and Montejoráez (2012) give excellent summaries. (The Martinez survey focuses specifically on tweets.)

In natural language systems, textual instances are often represented as vectors in a feature space. For example, if the space has only four features (f1, f2, f3, and f4), and each of these features is binary, that is they can have values 0 or 1, then an instance for which f1 is 0, f2 is 1, f3 is 1, and f4 is 0, can be represented by the vector $< 0, 1, 1, 0 >$. Training and test instances are converted into such feature vectors, which are in turn processed by the machine learning system. The number of features can often be as large as hundreds of thousands, and traditionally, these features have known meanings. For example, whether the instance has a particular word observed previously in the training data, whether the word is listed as a positive term in the sentiment lexicon, and so on. (Some work using uninterpretable features is described further ahead in the context of word embeddings.)

Word and character ngrams are widely used as features in a number of text classification problems, and it is not surprising to find that they are beneficial for valence classification as well. Features from manually and automatically created word–valence association lexicons, such as the General Inquirer (GI) (Stone, Dunphy, Smith, Ogilvie, & associates, 1966), the NRC Emotion Lexicon (Mohammad & Turney, 2010; Mohammad & Yang, 2011b), and SentiWordNet (SWN) (Esuli & Sebastiani, 2006), are aslo commonly used. Other features used by classification systems include those derived from parts of speech, punctuations (!, ???), word clusters, syntactic dependencies, negation terms (*no, not, never*), and word elongations (*hugggs, ahhhh*).

More recently, significant improvements classification accuracy have been obtained through *low-dimensional continuous representations* of instances and words (Collobert, Weston, Bottou, Karlen, Kavukcuoglu, & Kuksa, 2011; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Le & Mikolov, 2014). The phrase 'low-dimensional' refers to the notion that these vectors have only a few hundred dimension, and continuous refers to the real-valued nature of the dimensions (that is the dimension does not have just 0 or 1 values, but can have any real number value). These continuous word vectors, also called *embeddings*, are induced from a large corpus through neural networks (a particular kind of machine learning algorithm). The word vectors learned from the corpus are such that words that occur in similar contexts tend to be closer to each other in the low-dimensional space. However, unlike traditional feature vectors, these new dimensions are not directly interpretable, that is, it is not clear what any particular dimension signifies. First Socher, Perelygin, Wu, Chuang, Manning, Ng, and Potts (2013), and then Le and Mikolov (2014), obtained significant improvements in valence classification on a movie reviews dataset (Pang & Lee, 2008) using word embeddings. Work by Kalchbrenner, Grefenstette, and Blunsom (2014), Irsoy and Cardie (2014), Zhu, Sobhani, and Guo (2015), and others is further exploring the use recursive neural networks and word embeddings in sentiment analysis.

A number of shared task competitions on valence classification have been organized in recent years, including the 2013, 2014, and 2015 SemEval shared tasks titled *Sentiment Analysis in Twitter (SAT)*, the 2014 and 2015 SemEval shared tasks on *Aspect Based Sen-*

timent Analysis (ABSA), the 2015 SemEval shared task *Sentiment Analysis of Figurative Language in Twitter*, and the 2015 Kaggle competition *Sentiment Analysis on Movie Reviews*.[10] The SAT and ABSA tasks received submissions from more than 40 teams from universities, research labs, and companies across the world. The NRC-Canada system came first in the 2013 and 2014 SAT competitions (Mohammad, Kiritchenko, & Zhu, 2013b; Zhu et al., 2014b), and the 2014 ABSA competition (Kiritchenko et al., 2014a). The system is based on a supervised statistical text classification approach leveraging a variety of surface-form, semantic, and sentiment features. Notable, it used word and character ngrams, manually created and automatically generated sentiment lexicons, parts of speech, word clusters, and Twitter-specific encodings such as hashtags, creatively spelled words, and abbreviations (*yummeee, lol, etc*). The sentiment features were primarily derived from novel high-coverage tweet-specific sentiment lexicons. These lexicons were automatically generated from tweets with sentiment-word hashtags (such as *#great, #excellent*) and from tweets with emoticons (such as *:), :(*). (More details about these lexicons in in Section 5). Tang et al. (2014a) created a sentiment analysis system that came first in the 2014 SAT subtask on a tweets dataset. It replicated many of the same features used in the NRC-Canada system, and additionally used features drawn from word embeddings.

4.3 Automatically Detecting and Analyzing Emotions

Paul Ekman and others have developed theories on how some emotions are considered more basic than others (Ekman, 1992; Ekman & Friesen, 2003; Plutchik, 1980, 1991). These emotions are said to have ties to universal facial expressions and physiological processes such as increased heart rate and perspiration. However, not everybody agrees on which set of emotions are the most basic. Ekman (1992), Plutchik (1980), Parrot (2001), Frijda (1988), Shaver, Schwartz, Kirson, and O'connor (1987), and others proposed different sets of basic emotions. Even more controversially, the very theory of basic emotions has been challenged in recent work (Barrett, 2006; Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; De Leersnyder, Boiger, & Mesquita, 2015). Nonetheless, much of the efforts in automatic detection of emotions in text has focused on the handful of proposed basic emotions. Recall that labeled training data is a crucial resource required for building supervised machine learning systems. Compiling datasets with tens of thousands of instances annotated for emotion is expensive in terms of time and money. Each instance must be annotated by more than one person (usually three to five) to determine how much people agree with each other on emotion annotations. Asking for labels from a large set of emotions increases cognitive load on the annotator. Asking annotators to label the data one emotion at a time, repeated for a large number of emotions, increases the cost of annotation. Thus, focusing on a small number of emotions has the benefit of keeping costs down. On the other hand, work focused on a small set of emotions means that there are fewer resources and systems that can handle non-basic emotions. Further, different emotions may be more relevant for different cultures (De Leersnyder et al., 2015).

10. http://aclweb.org/aclwiki/index.php?title=SemEval_Portal
http://alt.qcri.org/semeval2015/task10/
http://alt.qcri.org/semeval2015/task12/
http://alt.qcri.org/semeval2015/task11/
http://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

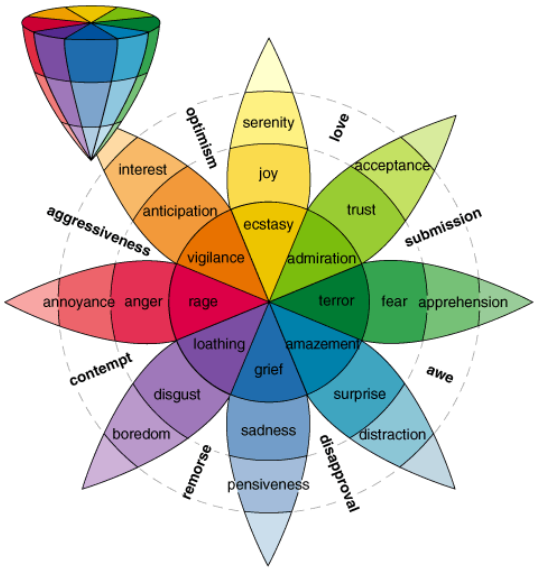

Figure 1: Set of eight basic emotions proposed in Plutchik (1980).

Below we summarize work on compiling textual datasets labeled with emotions and automatic methods for detecting emotions in text. We group the work by the emotion categories addressed.

- **Work on Ekman's Six**: Paul Ekman's set of basic emotions includes: joy, sadness, anger, fear, disgust, and surprise. Since the writing style and vocabulary in different sources, such as chat messages, blog posts, and news paper articles, can be very different, automatic systems that cater to specific domains are more accurate when trained on data from the target domain. Holzman and Pottenger (2003) annotated 1201 chat messages for the Ekman's six emotions as well as for irony and neutral classes. Alm, Roth, and Sproat (2005) annotated 22 Grimm fairy tales (1580 sentences) for Ekman emotions.[11] Strapparava and Mihalcea (2007) annotated news paper headlines with intensity scores for each of the Ekman emotions, referred to as the *Text Affect Dataset*.[12] Aman and Szpakowicz (2007) annotated blog posts with the Ekman emotions. These datasets have been used by its creators to develop supervised machine learning algorithms to classify instances into one of the six Ekman classes (or neutral). Other work that developed classification systems on these datasets includes work by Chaffar and Inkpen (2011) (on Text Affect and datasets created by Alm and Aman); Mohammad (2012) (on Text Affect and Aman datasets); and Kirange et al. (2013) (on Text Affect).

- **Work on Plutchik's Eight**: Robert Plutchik's set of basic emotions includes Ekman's six as well as trust and anticipation. Figure 1 shows how Plutchik arranges

11. https://dl.dropboxusercontent.com/u/585585/RESOURCEWEBSITE1/index.html
12. http://web.eecs.umich.edu/~mihalcea/downloads.html#affective

these emotions on a wheel such that opposite emotions appear diametrically opposite to each other. Words closer to the center have higher intensity than those that are farther. Plutchik also hypothesized how some secondary emotions can be seen as combinations of some of the basic (primary) emotions, for example, optimism as the combination of joy and anticipation. See Plutchik (1991, 2001) for details about his taxonomy of emotions created by primary, secondary, and tertiary emotions.

Brooks, Kuksenok, Torkildson, Perry, Robinson, Scott, Anicello, Zukowski, Harris, and Aragon (2013) annotated 27,344 chat messages between thirty astrophysics collaborators with 40 affect categories inspired by Plutchik's taxonomy emotions. Mohammad (2012a) polled the Twitter API for tweets that have hashtag words such as *#anger* and *#sadness* corresponding to the eight Plutchik emotions.[13] He showed that these hashtag words act as good labels for the rest of the tweets, and that this labeled dataset is just as good as the set explicitly annotated for emotions for emotion classification. Such an approach to machine learning from pseudo-labeled data is referred to as *distant supervision*. Suttles and Ide (2013) used a similar distant supervision technique and collected tweets with emoticons, emoji, and hashtag words corresponding to the Plutchik emotions. They developed an algorithm for binary classification of tweets along the four opposing Plutchik dimensions. Kunneman, Liebrecht, and van den Bosch (2014) studied the extent to which hashtag words in tweets are predictive of the affectual state of the rest of the tweet. They found that hashtags can vary significantly in this regard—some hashtags are strong indicators of the corresponding emotion whereas others are not. Thus hashtag words must be chosen carefully when employing them for distant supervision.

- **Work on Other Small Sets of Emotions**: The ISEAR Project asked 3000 student respondents to report situations in which they had experienced joy, fear, anger, sadness, disgust, shame, or guilt.[14] Thomas et al. (2014) applied supervised machine learning techniques on the ISEAR dataset for 7-way emotion classification. Neviarouskaya et al. (2009) collected 1000 sentences from the Experience Project webpage and manually annotated them for fourteen affectual categories.[15] Experience Project is a portal where users share their life experiences. These shared texts are usually rife with emotion.

  Pearl and Steyvers (2010) developed an online Game With a Purpose (GWAP) where participants were asked to generate labels for politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief. Then other participants would label these messages to determine whether multiple people agree that the message belongs to the same category.

  Bollen et al. (2009) analyzed 9,664,952 tweets posted in the second half of 2008 using Profile of Mood States (POMS) (McNair, Lorr, & Droppleman, 1989). POMS is a psychometric instrument that measures the mood states of tension, depression, anger, vigor, fatigue, and confusion.

13. http://saifmohammad.com/WebPages/lexicons.html
14. http://emotion-research.net/toolbox/toolboxdatabase.2006-10-13.2581092615
15. www.experienceproject.com

Wang et al. (2012) compiled a set of 2.5 milion tweets with emotion-related hashtags using the distant supervision technique. The emotion-related hashtags correspond to seven emotion categories: joy, sadness, anger, love, fear, thankfulness, and surprise. They also developed a machine learning algorithm to classify tweets into these seven emotion categories and found the most useful features to be unigrams, bigrams, sentiment and emotion lexicons (LIWC, MPQA, WordNet Affect), and part of speech.

- **Work on Emotion-Labeled Datsets in Languages Other than English:** Wang (2014) annotated Chinese news and blog posts with the Ekman emotions. Wang also translated Alm's fairy tales dataset into Chinese. Quan and Ren (2009) created a blog emotion corpus in Chinese called the *Ren-CECps Corpus*. The sentences in this corpus are annotated with eight emotions: expectation, joy, love, surprise, anxiety, sorrow, anger, and hate. Sentences not associated with any of these eight categories are marked as neutral. The corpus has 1,487 documents, 11,255 paragraphs, 35,096 sentences, and 878,164 Chinese words.

  The 2013 Chinese Microblog Sentiment Analysis Evaluation (CMSAE) compiled a dataset of posts from Sina Weibo (a popular Chinese microblogging service) annotated with seven emotions: anger, disgust, fear, happiness, like, sadness and surprise.[16] If a post has no emotion, then it is labeled as *none*. The training set contains 4000 instances (13252 sentences). The test dataset contains 10000 instances (32185 sentences). Wen and Wan (2014) developed model to detect emotions in this corpus by combining lexcion-based and SVM-based methods.

  Sun, Quan, Kang, Zhang, and Ren (2014) created a Japanese customer reviews corpus with the same eight emotions used in the Chinese *Ren-CECps Corpus*: expectation, joy, love, surprise, anxiety, sorrow, anger, and hate. The annotated corpus has 3,264 sentences. Each adverb and sentence was manually annotated for association with the eight emotions and also the degree of emotion intensity (0.1 to 1.0). They also created an adverb emotion lexicon which contains 687 adverbs and their associations with the eight emotions.

- **Large Sets of Emotions:** Distant supervision techniques proposed in Mohammad (2012a), Purver and Battersby (2012), Wang et al. (2012), Suttles and Ide (2013), and others have opened up the critical bottleneck of creating instances labeled with emotion categories. Thus, now, labeled data can be created for any emotion for which there are sufficient number of tweets that have the emotion word as a hashtag. Mohammad and Kiritchenko (2014) collected tweets with hashtags corresponding to around 500 emotion words as well as positive and negative valence. They used these tweets to identify words associated with each of the 500 emotion categories, which were in turn used as features in a task of automatically determining personality traits from stream-of-consciousness essays. They show that using features from 500 emotion categories significantly improved performance over using features from just the Ekman emotions.

As seen above, there are a number of datasets where sentences are manually labeled for emotions. They have helped take the field forward by allowing researchers to understand

16. http://tcci.ccf.org.cn/conference/2013/pages/page04_eva.html

what it means to annotate text for affect categories, and by helping develop supervised machine learning emotion classifiers. Even though features from word ngrams, part of speech, and term–emotion association lexicons are commonly used, new features based on continuous word representations are now leading to more accurate emotion classification, just as in valence classification. Yet, a number of questions remain unexplored. For example, can the large amounts of textual data (and self-labeled data as in tweets with emotion-word hashtags) be used to infer if indeed there are some emotions that are more basic than others or whether there is a taxonomy of emotions? Are some emotions indeed combinations of other emotions (optimism as the combination of joy and anticipation)? Can data labeled for certain emotions be useful in detecting certain other emotions? Can automatic systems that detect degree of valence, arousal, and dominance be used to infer emotions such as joy, sadness, fear, etc.? And so on.

There is new work on developing word representations not only from text corpora but also from collections of images that the words are associated with (Kiela & Bottou, 2014; Hill & Korhonen, 2014; Lazaridou, Bruni, & Baroni, 2014). (An image and a word can be considered associated if the caption for the image has the word, the image is a representation of the concept the word refers to, etc.) This bridges the gap between text and vision, allowing the exploration of new applications such as automatic image captioning (Karpathy, Joulin, & Li, 2014; Kiros, Salakhutdinov, & Zemel, 2014). Future work can explore how emotions should influence such multi-modal word representations (text–vision, text–audio, etc.) so as to obtain even better representations for emotions. Such multi-modal representations of emotions will be useful in tasks such as captioning images or audio for emotions and even generating text that is affectually suitable for a give image or audio sequence.

5. Capturing Term–Sentiment Associations

The same word can convey different sentiment in different contexts. For example, the word *unpredictable* is negative in the context of automobile steering, but positive in the context of a movie script. Nonetheless, many words have a tendency to convey the same sentiment in a large majority of the contexts they occur in. For example, *excellent* and *cake* are positive in most usages whereas *death* and *depression* are negative in most usages. These majority associations are referred to as *prior associations*. Sentiment analysis systems benefit from knowing these prior associations of words and phrases. Thus, lists of term–sentiment associations have been created by manual annotation. These resources tend to be small in coverage because manual annotation is expensive and the number of words and phrases for a language can run into hundreds of thousands. This has led to the development of automatic methods that extract large lists of term–sentiment associations from text corpora using manually created lists as seeds. We describe work on manually creating and automatically generating term–sentiment associations in the sub-sections below.

5.1 Manually generated term-sentiment association lexicons

One of the earliest works exploring term–sentiment associations was by Osgood et al. (1957) who in their book, *Measurement of Meaning*, describe an experiment in which they asked respondents to state where various words lie within several semantic dimensions. These semantic dimensions were formed by bipolar adjectives such as *adequate–inadequate*, *good–evil*,

and *valuable–worthless*. Factor analysis of these ratings showed that the three dimensions across which people judge a word most (in decreasing order) are evaluation (*good–bad*), potency (*strong–weak*), and activity (*active–passive*). Evaluativeness can be thought of as the same dimension as valence (positive–negative). The General Inquirer (GI) lists words associated with various semantic categories including evaluativeness for about 3,600 terms (Stone et al., 1966). These include about 1500 words from the Osgood study. The MPQA Subjectivity Lexicon, which draws from the General Inquirer and other sources, has valence labels for about 8,000 words (Wilson, Wiebe, & Hoffmann, 2005). The MPQA lexicon categorizes terms into strongly positive, weakly positive, strongly negative and weakly negative. Hu and Liu (2004) manually labeled about 6,800 words and used them for detecting sentiment of customer reviews. The Affective Norms for English Words (ANEW) provides valence, arousal, and dominance ratings for 1034 English words (Bradley & Lang, 1999).[17] AFINN (Nielsen, 2011) has valence ratings for 2477 English words. The ratings range from -5 (most negative) to +5 (most positive) in steps of 1.[18]

A new trend in creating large amounts of human-annotated data is *crowdsourcing*. Crowdsourcing involves breaking down a large task into small independently solvable units, distributing the units through the Internet or some other means, and getting a large number of people to solve or annotate the units. The requester specifies the compensation that will be paid for solving each unit. In this scenario, the annotators are usually not known to the requester and usually do not all have the same academic qualifications. Natural language tasks are particularly well-suited for crowdsourcing because even though computers find it difficult to understand language, native speakers of a language do not usually need extensive training to provide useful annotations such as whether a word is associated with positive sentiment. Amazon Mechanical Turk and CrowdFlower are two commonly used crowdsourcing platforms.[19] They allow for large scale annotations, quickly and inexpensively. However, one must define the task carefully to obtain annotations of high quality. Checks must be placed to ensure that random and erroneous annotations are discouraged, rejected, and re-annotated.

The NRC Emotion Lexicon (Mohammad & Turney, 2010, 2012) was created by crowdsourcing, and it has associations towards positive and negative sentiment as well as the eight Plutchik emotions—joy, sadness, fear, anger, anticipation, trust, surprise, and disgust (Plutchik, 1962, 1980). Respondents were biased to individual senses of a word by priming the target word with another word relevant to a particular sense. They were asked whether a term is strongly associated, moderately associated, weakly associated, or not associated with the target. The lexicon has valence and emotion associations for about 25,000 word senses. A word-level version of the lexicon created by taking the union of associations of all the senses of a word. It has valence and emotion labels for about 14,000 words. This version also collapsed the strong and moderate associations into one *associated* category and the weak and no associations into one *not associated* category. The listing of words in the *not associated* category is useful to distinguish between words for which we know they are not associated with the Plutchik emotion categories and words for which we do not know

17. http://csea.phhp.ufl.edu/media/anewmessage.html
18. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
19. https://www.mturk.com/mturk/welcome
    http://www.crowdflower.com

whether they are associated with any of the Plutchik emotions. The terms that are not marked to be associated with any valence are also useful as seed words that can be used to generate larger lists of neutral expressions (Bravo-Marquez, Frank, & Pfahringer, 2015).

Warriner, Kuperman, and Brysbaert (2013) created a crowdsourced lexicon with valence, arousal, and dominance annotations for 13,915 terms. A subset of entries in the lexicon correspond to words in ANEW. For these terms, the authors show that the valence annotations obtained by crowdsourcing have high correlation with the ratings in ANEW. The correlation is lower for arousal and dominance, but that is not very different from other independent comparisons of arousal and dominance ratings.

All of the lexicons described so far in this section have been widely used in natural language processing research. However, none of them provide real-valued scores indicating degrees of association of terms with valence or emotion categories.

5.1.1 Real-valued Sentiment Scores from manual Annotations

Words have varying degree of associations with sentiment categories. This is true not just for comparative and superlative adjectives and adverbs (for example, *worst* is more negative than *bad*) but also for other syntactic categories. For example, most people will agree that *succeed* is more positive (or less negative) than *improve*, and *fail* is more negative (or less positive) than *deteriorate*. Downstream applications benefit from knowing not only whether a word or phrase in positive or negative (or associated with some emotion category), but also from knowing the strength of association. However, for people, assigning a score indicating the degree of sentiment is not natural. Different people may assign different scores to the same target item, and it is hard for even the same annotator to remain consistent when annotating a large number of items. In contrast, it is easier for annotators to determine whether one word is more positive (or more negative) than the other. However, the latter requires a much larger number of annotations than the former (in the order of N^2, where N is the number of items to be annotated).

An annotation scheme that retains the comparative aspect of annotation while still requiring only a small number of annotations comes from survey analysis techniques and is called *maximum difference scaling (MaxDiff)* or *best-worst scaling* (Louviere, 1991).

The annotator is presented with four terms and asked which word is the most positive and which is the least positive. By answering just these two questions five out of the six inequalities are known. If the respondent says that A is most positive and D is least positive, then:

$$A > B, A > C, A > D, B > D, C > D$$

Each of these MaxDiff questions can be presented to multiple annotators. The responses to the MaxDiff questions can then be easily translated into a ranking of all the terms and also a real-valued score for all the terms (Orme, 2009). If two words have very different degrees of association (for example, $A >> D$), then A will be chosen as most positive much more often than D and D will be chosen as least positive much more often than A. This will eventually lead to a ranked list such that A and D are significantly farther apart, and their real-valued association scores are also significantly different. On the other hand, if two words have similar degrees of association with positive sentiment (for example, A and B), then it is possible that for MaxDiff questions having both A and B, some annotators

will choose A as most positive, and some will choose B as most positive. Further, both A and B will be chosen as most positive (or most negative) a similar number of times. This will result in a list such that A and B are ranked close to each other and their real-valued association scores will also be close in value.

MaxDiff was used for obtaining annotations of relation similarity of pairs of items by (Jurgens, Mohammad, Turney, & Holyoak, 2012) in a SemEval-2012 shared task. Kiritchenko et al. (2014b) used the MaxDiff method to create a dataset of 1500 Twitter terms with real-valued sentiment association scores. They also conducted an experiment to determine the reliability of the sentiment scores by randomly dividing the responses into two groups and comparing the sentiment scores. obtained from the two groups. On average, the scores differed only by 0.04, showing good reliability.

Real-valued valence association scores obtained through MaxDiff Annotations were used in subtask E of the 2015 SemEval Task *Sentiment Analysis in Twitter* (Rosenthal, Nakov, Kiritchenko, Mohammad, Ritter, & Stoyanov, 2015) to evaluate automatically generated Twitter-specific valence lexicons. Datasets created with the same approach will be used in a 2016 Task *Determining sentiment intensity of English and Arabic phrases* to evaluate both English and Arabic automatically generated sentiment lexicons.

5.2 Automatically generated term–sentiment association lexicons

Automatic, statistical, methods for capturing word–sentiment associations can quickly learn associations for hundreds of thousands words, and even for sequences of words. They can also learn associations that are relevant to a particular domain. For example, when the algorithm is applied on a text of movie reviews, the system can learn that *unpredictable* is a positive term in this domain (as in *unpredictable story line*), but when applied to auto reviews, the system can learn that *unpredictable* is a negative term (as in *unpredictable steering*).

Hatzivassiloglou and McKeown (1997) proposed an algorithm that uses word usage patterns to generate a graph with adjectives as nodes. An edge between two nodes indicates either that the two adjectives have the same or opposite valence. A clustering algorithm then partitions the graph into two subgraphs such that the nodes in a subgraph have the same valence. They used this method to create a lexicon of positive and negative words.

Turney and Littman (2003) proposed a minimally supervised algorithm to calculate the valence of a word by determining if its tendency to co-occur with a small set of positive seed words is greater than its tendency to co-occur with a small set of negative seed words. SentiWordNet (SWN) was created using supervised classifiers as well as manual annotation (Esuli & Sebastiani, 2006). Mohammad, Dunne, and Dorr (2009) automatically generated a sentiment lexicon of more than 60,000 words from a thesaurus.

Mohammad et al. (2013a) employed the Turney method to generate a lexicon (Hashtag Sentiment Lexicon) from tweets with certain sentiment-bearing seed-word hashtags such as (*#excellent, #good, #terrible,* and so on) and another lexicon (Hashtag Sentiment Lexicon) from tweets with emoticons.[20] Since the lexicons themselves are generated from tweets, they even have entries for the creatively spelled words (e.g. *happpeee*), slang (e.g. *bling*), abbreviations (e.g. *lol*), and even hashtags and conjoined words (e.g. *#loveumom*). Kir-

20. http://www.purl.com/net/lexicons

itchenko et al. (2014b) proposed a method to create separate lexicons for words found in negated context and those found in affirmative context; the idea being that the same word contributes to sentiment differently depending on whether it is negated or not. These lexicons contain sentiment associations for hundreds of thousands of unigrams and bigrams. However, they do not explicitly handle combinations of terms with modals, degree adverbs, and intensifiers.

Other recent work on valence lexicons includes Tang, Wei, Qin, Zhou, and Liu (2014b), Chetviorkin, Moscow, and Loukachevitch (2014), Chetviorkin et al. (2014), Makki, Brooks, and Milios (2014). Tang et al. (2014b) proposed a method to determine large valence association lexicons from tweets using a neural network architecture and a continuous representation approach. They evaluate their approach by measuring usefulness in tweet valence classification tasks. Chetviorkin et al. (2014) proposed a method for constructing domain-specific valence lexicons.

6. Modeling the impact of sentiment modifiers

Negation, modality, degree adverbs and other modifiers impact the sentiment of the term or phrase they modify.

6.1 Negation

Morante and Sporleder (2012) define negation to be "a grammatical category that allows the changing of the truth value of a proposition". Negation is often expressed through the use of negative signals or negator words such as *not* and *never*, and it can significantly affect the sentiment of its scope. Understanding the impact of negation on sentiment improves automatic detection of sentiment.

Automatic negation handling involves identifying a negation word such as *not*, determining the scope of negation (which words are affected by the negation word), and finally appropriately capturing the impact of the negation. (See work by Jia, Yu, and Meng (2009), Wiegand, Balahur, Roth, Klakow, and Montoyo (2010), Lapponi, Read, and Ovrelid (2012) for detailed analyses of negation handling.) Traditionally, the negation word is determined from a small hand-crafted list (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). The scope of negation is often assumed to begin from the word following the negation word until the next punctuation mark or the end of the sentence (Polanyi & Zaenen, 2004; Kennedy & Inkpen, 2005). More sophisticated methods to detect the scope of negation through semantic parsing have also been proposed (Li, Zhou, Wang, & Zhu, 2010).

Earlier works on negation handling employ simple heuristics such as flipping the polarity of the words in a negator's scope (Kennedy & Inkpen, 2005; Choi & Cardie, 2008) or changing the degree of sentiment of the modified word by a fixed constant (Taboada et al., 2011). Zhu, Guo, Mohammad, and Kiritchenko (2014a) show that these simple heuristics fail to capture the true impact of negators on the words in their scope. They show that negators tend to often make positive words negative (albeit with lower intensity) and make negative words less negative (not positive). Zhu et al. also propose certain embeddings-based recursive neural network models to capture the impact of negators more precisely. As mentioned earlier, Kiritchenko et al. (2014b) capture the impact of negation by creating separate sentiment lexicons for words seen in affirmative context and those seen in negated

contexts. These lexicons are generated using co-occurrence statistics of terms in affirmative context with sentiment signifiers such as emoticons and seed hashtags (such as *#great, #horrible*), and separately for terms in negated contexts with sentiment signifiers. They use a hand-chosen list of negators and determine scope to be starting from the negator and ending at the first punctuation (or end of sentence).

6.2 Degree Adverbs, Intensifiers, Modals

Degree adverbs such as *barely, moderately*, and *slightly* quantify the extent or amount of the predicate. Intensifiers such as *too* and *very* are modifiers that do not change the propositional content (or truth value) of the predicate they modify, but they add to the emotionality. However, even linguists are hard pressed to give out comprehensive lists of degree adverbs and intensifiers. Additionally, the boundaries between degree adverbs and intensifiers can sometimes be blurred, and so it is not surprising that the terms are occasionally used interchangeably. Impacting propositional content or not, both degree adverbs and intensifiers impact the sentiment of the predicate, and there is some work in exploring this interaction (Zhang, Zeng, Xu, Xin, Mao, & Wang, 2008; Wang & Wang, 2012; Xu, Wong, Lu, Xia, & Li, 2008; Lu & Tsou, 2010; Taboada, Voll, & Brooke, 2008). Most of this work focuses on identifying sentiment words by bootstrapping over patterns involving degree adverbs and intensifiers. Thus several areas remain unexplored, such as identifying patterns and regularities in how different kinds of degree adverbs and intensifiers impact sentiment, ranking degree adverbs and intensifiers in terms of how they impact sentiment, and determining when (in what contexts) the same modifier will impact sentiment differently than its usual behavior.

Modals are a kind of auxiliary verb used to convey the degree of confidence, permission, or obligation. Examples include *can, could, may, might, must, will, would, shall,* and *should.* The sentiment of the combination of the modal and an expression can be different from the sentiment of the expression alone. For example, *cannot work* is less positive than *work* or *will work* (*cannot* and *will* are modals). Thus handling modality appropriately can greatly improve automatic sentiment analysis systems.

7. Sentiment in figurative and metaphoric language

There is growing interest in detecting figurative language, especially irony and sarcasm (Carvalho, Sarmento, Silva, & De Oliveira, 2009; Reyes, Rosso, & Veale, 2013; Veale & Hao, 2010; Filatova, 2012; González-Ibánez, Muresan, & Wacholder, 2011). In 2015, a SemEval shared task was organized on detecting sentiment in tweets rich in metaphor and irony (Task 11).[21] Participants were asked to determine the degree of sentiment for each tweet where the score is a real number in the range from -5 (most negative) to +5 (most positive). One of the characteristics of the data is that most of the tweets are negative; thereby suggesting that ironic tweets are largely negative. The SemEval 2014 shared task Sentiment Analysis in Twitter (Rosenthal et al., 2014) had a separate test set involving sarcastic tweets. Participants were asked *not* to train their system on sarcastic tweets, but rather apply their regular sentiment system on this new test set; the goal was to determine

21. The proceedings will be released later in 2015.

performance of regular sentiment systems on sarcastic tweets. It was observed that the performances dropped by about 25 to 70 percent, thereby showing that systems must be adjusted if they are to be applied to sarcastic tweets. We found little to no work exploring automatic sentiment detection in hyperbole, understatement, rhetorical questions, and other creative uses of language.

8. Multilingual Sentiment Analysis

A large proportion of research in sentiment analysis has focused on English. Thus there are fewer resources (sentiment lexicons, annotated corpora, etc) for other languages than in English. This means that automatic sentiment analysis systems in other languages tend to be less accurate than their English counterpart. Thus work on multilingual sentiment analysis has mainly addressed mapping sentiment resources from English into morphologically complex languages. Mihalcea, Banea, and Wiebe (2007) use English resources to automatically generate a Romanian subjectivity lexicon using an English–Romanian dictionary. The generated lexicon is then used to classify Romanian text. Wan (2008) translated Chinese customer reviews to English using a machine translation system. The translated reviews are then annotated using rule-based system that uses English lexicons. A higher accuracy is achieved when using ensemble methods and combining knowledge from Chinese and English resources.

Often companies, organizations, and governments need to convey information in many languages. With capabilities in automatic translation improving every year, it is tempting to produce text in one language (say English) and translate into other languages using an automatic system. One might then want to manually inspect the translations to correct for errors. However, translations may not always preserve the sentiment in the source text. Balahur and Turchi (2014) conducted a study to assess the performance of sentiment analysis techniques on machine-translated texts. Opinion-bearing English phrases from the New York Times Text (2002–2005) corpus were split into training and test datasets. An English sentiment analysis system was trained on the training dataset and its prediction accuracy on the test set was found to be about 68%. Next, the training and test datasets were automatically translated into German, Spanish, and French using publicly available machine-translation engines (Google, Bing, and Moses). The translated test sets were then manually corrected for errors. Then for German, Spanish, and French, a sentiment analysis system was trained on the translated training set for that language and tested on the translated-and-corrected test set. The authors observe that these German, Spanish, and French sentiment analysis systems obtain accuracies in the low sixties (and thus not very much lower than 68%). However, the languages explored in this study are linguistically close to each other.

Salameh, Mohammad, and Kiritchenko (2015) and Mohammad, Salameh, and Kiritchenko (2015) conducted experiments to determine loss in sentiment predictability when they translate Arabic social media posts into English, manually and automatically. As a benchmark, they use manually determined sentiment of the Arabic text. They show that an English sentiment analysis system has only a slightly lower accuracy on the English translation of Arabic text as compared to the accuracy of an Arabic sentiment analysis system on the original (untranslated) Arabic text. This does not imply that translation

does not have an effect on valence. On the contrary, Salameh et al. (2015) and Mohammad et al. (2015) show that even with manual translations of text, cultural differences can lead to significantly different valence associations between speakers of the two languages. They also showed that automatic Arabic translations of English valence lexicons improve accuracies of an Arabic sentiment analysis system. The translated lexica and corpora are made freely available.[22] The experiments with automatic translations and automatic sentiment analysis systems show that in languages where a strong sentiment analysis system does not exist, using an English sentiment analysis system on English translations of text from that language is a viable option.

Some of the areas less explored in the realm of multilingual sentiment analysis include: how to translate text so as to preserve the degree of sentiment in the source text; how sentiment modifiers such as negators and modals differ in function across languages; understanding how automatic translations differ from manual translations in terms of sentiment; and how to translate figurative language without losing its affectual gist.

9. Summary and Future Directions

This chapter summarized the diverse landscape of problems and applications associated with automatic sentiment analysis. We outlined key challenges for automatic systems, as well as the algorithms, features, and datasets used in sentiment analysis. We described several manual and automatic approaches to creating valence- and emotion-association lexicons. We also described work on sentence-level sentiment analysis. We discussed preliminary approaches to handle sentiment modification by negators and modals, detecting sentiment in figurative and metaphoric language, as well as cross-lingual sentiment analysis—these are areas where we expect to see significantly more work in the near future. Other promising areas of future work include: understanding the relationships between emotions, multimodal affect analysis (involving not just text but also speech, vision, physiological sensors, etc), and applying emotion detection to new applications.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of Language in Social Media*, pp. 30–38, Portland, Oregon.

Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012). Textual and contextual patterns for sentiment analysis over microblogs. In *Proceedings of the 21st WWW Companion*, pp. 453–454, New York, NY, USA.

Alm, C. O. (2008). *Affect in text and speech*. ProQuest.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Joint Conference on HLT–EMNLP*, Vancouver, Canada.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, Vol. 4629 of *Lecture Notes in Computer Science*, pp. 196–205.

22. Lexicons and corpora for Arabic sentiment analysis: www.purl.org/net/ArabicSA

Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155–177.

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., & Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pp. 1–9.

Avello, D. G. (2012). "i wanted to predict elections with twitter and all i got was this lousy paper" – a balanced survey on election prediction using twitter data. *arXiv*, *1204.6441*.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley framenet project. In *Proceedings of ACL*, pp. 86–90, Stroudsburg, PA.

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. In *Proceedings of WASSA'12*, pp. 11–18, Jeju, Republic of Korea.

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

Barrett, L. F. (2006). Are emotions natural kinds?. *Perspectives on psychological science*, *1*(1), 28–58.

Bellegarda, J. (2010). Emotion analysis using latent affective folding and embedding. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

Bermingham, A., & Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. *Psychology*, 2–10.

Bock, R., Gluge, S., Wendemuth, A., Limbrecht, K., Walter, S., Hrabal, D., & Traue, H. C. (2012). Intraindividual and interindividual multimodal emotion analyses in human-machine-interaction. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2012 IEEE International Multi-Disciplinary Conference on*, pp. 59–64. IEEE.

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*.

Boucouvalas, A. C. (2002). Real time text-to-emotion engine for expressive internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication*, *5*, 305–318.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Tech. rep., Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Bravo-Marquez, F., Frank, E., & Pfahringer, B. (2015). Positive, negative, or neutral: Learning an expanded opinion lexicon from emoticon-annotated tweets..

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 562–570.

Brooks, M., Kuksenok, K., Torkildson, M. K., Perry, D., Robinson, J. J., Scott, T. J., Anicello, O., Zukowski, A., Harris, P., & Aragon, C. R. (2013). Statistical affect detection in collaborative chat. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 317–328. ACM.

Carvalho, P., Sarmento, L., Silva, M. J., & De Oliveira, E. (2009). Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pp. 53–56. ACM.

Chaffar, S., & Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. In *Canadian Conference on AI*, pp. 62–67.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 71–80. IEEE.

Cherry, C., Mohammad, S. M., & De Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical informatics insights*, *5*(Suppl 1), 147.

Chetviorkin, I., Moscow, L. G., & Loukachevitch, N. (2014). Two-step model for sentiment lexicon extraction from twitter streams. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 67–72, Baltimore, MD, USA.

Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 793–801, Honolulu, Hawaii.

Choudhury, M. D., Counts, S., & Gamon, M. (2012). Not all moods are created equal! exploring human emotional states in social media. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493–2537.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonc, B., Flammini, A., & Menczer, F. (2011a). Political polarization on Twitter. *Networks*, *133*(26), 89–96.

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011b). Predicting the political alignment of Twitter users. In *IEEE Third International Conference on Privacy Security Risk and Trust and IEEE Third International Conference on Social Computing*, pp. 192–199. IEEE.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pp. 693–696. Springer.

Das, A., & Bandyopadhyay, S. (2009). Subjectivity detection in english and bengali: A crf-based approach. *Proceeding of ICON*.

Davis, H., & Mohammad, S. (2014). Generating music from literature. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 1–10, Gothenburg, Sweden.

De Leersnyder, J., Boiger, M., & Mesquita, B. (2015). Cultural differences in emotions. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*.

Dogan, H. (2012). Emotion, confidence, perception and expectation case of mathematics. *International Journal of Science and Mathematics Education*, *10*(1), 49–69.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., et al. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*.

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

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC '06, pp. 417–422.

Fang, R.-Y., Chen, B.-W., Wang, J.-F., & Wu, C.-H. (2011). Emotion detection based on concept inference and spoken sentence analysis for customer service. In *Twelfth Annual Conference of the International Speech Communication Association*.

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, pp. 392–398.

Fillmore, C. (1982). Frame semantics. *Linguistics in the morning calm*, 111–137.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, Vol. 280, pp. 20–32.

Francisco, V., & Gervás, P. (2006). Automated mark up of affective information in english texts. In Sojka, P., Kopecek, I., & Pala, K. (Eds.), *Text, Speech and Dialogue*, Vol. 4188 of *Lecture Notes in Computer Science*, pp. 375–382. Springer Berlin / Heidelberg.

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

Genereux, M., & Evans, R. P. (2006). Distinguishing affective states in weblogs. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 27–29, Stanford, California.

Ghazi, D., Inkpen, D., & Szpakowicz, S. (2015). Detecting emotion stimuli in emotion-bearing sentences. In *Proceedings of the 2015 Conference on Intelligent Text Processing and Computational Linguistics*.

Gobron, S., Ahn, J., Paltoglou, G., Thelwall, M., & Thalmann, D. (2010). From sentence to emotion: a real-time three-dimensional graphics metaphor of emotions extracted from text. *The Visual Computer*, *26*(6-8), 505–519.

Golbeck, J., & Hansen, D. (2011). Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 1105–1108, New York, NY. ACM.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the ACL*, pp. 581–586. Association for Computational Linguistics.

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P., Robins, R. W., & Yan, T. (2014). Gender differences in narcissism: A meta-analytic review...

Gupta, N., Gilbert, M., & Fabbrizio, G. D. (2013). Emotion detection in email customer care. *Computational Intelligence*, *29*(3), 489–505.

Hartner, M. (2013). The lingering after-effects in the readers mind – an investigation into the affective dimension of literary reading. *Journal of Literary Theory Online*.

Hasan, K. S., & Ng, V. (2013). Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1348–1356.

Hasan, M., Rundensteiner, E., & Agu, E. (2014). Emotex: Detecting emotions in twitter messages..

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference of European Chapter of the Association for Computational Linguistics*, pp. 174–181, Madrid, Spain.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pp. 299–305. Association for Computational Linguistics.

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably cant see what i mean. *Proceedings of EMNLP. ACL*.

Holzman, L. E., & Pottenger, W. M. (2003). Classification of emotions in internet chat: An application of machine learning using speech phonemes. Tech. rep., Leigh University.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, New York, NY, USA. ACM.

Irsoy, O., & Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems*, pp. 2096–2104.

Jia, L., Yu, C., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 1827–1830, New York, NY, USA. ACM.

John, D., Boucouvalas, A. C., & Xu, Z. (2006). Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED international*

conference on Internet and multimedia systems and applications, pp. 183–188, Anaheim, CA. ACTA Press.

Johnsen, J.-A. K., Vambheim, S. M., Wynn, R., & Wangberg, S. C. (2014). Language of motivation and emotion in an internet support group for smoking cessation: explorative use of automated content analysis to measure regulatory focus. *Psychology research and behavior management*, *7*, 19.

Jurgens, D., Mohammad, S. M., Turney, P., & Holyoak, K. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pp. 356–364, Montréal, Canada.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Karpathy, A., Joulin, A., & Li, F. F. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pp. 1889–1897.

Kennedy, A., & Inkpen, D. (2005). Sentiment classification of movie and product reviews using contextual valence shifters. In *Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*, Ottawa, Ontario, Canada.

Kessler, J. S., & Nicolov, N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In *3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, USA.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, Vol. 2014.

Kilgarriff, A. (1997). I dont believe in word senses. *Computers and the Humanities*, *31*(2), 91–113.

Kim, E., Gilbert, S., Edwards, M. J., & Graeff, E. (2009). Detecting sadness in 140 characters: Sentiment analysis of mourning Michael Jackson on twitter..

Kirange, D., et al. (2013). Emotion classification of news headlines using svm. *Asian Journal of Computer Science & Information Technology*, *2*(5).

Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. M. (2014a). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014b). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kleres, J. (2011). Emotions and narrative analysis: A methodological approach. *Journal for the theory of social behaviour*, *41*(2), 182–202.

Kramer, A. D. (2012). The spread of emotion via facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 767–770. ACM.

Ku, L.-W., Liang, Y.-T., & Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 100107.

Kunneman, F., Liebrecht, C., & van den Bosch, A. (2014). The (un) predictability of emotional hashtags in twitter. In *Proceedings of the 5th Workshop on Language Analysis for Social Media*, pp. 26–34, Gothenburg, Sweden.

Lampos, V., Preotiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from social media. In *Proc 51st Annual Meeting of the Association for Computational Linguistics*, pp. 993–1003.

Lapponi, E., Read, J., & Ovrelid, L. (2012). Representing and resolving negation for sentiment analysis. In Vreeken, J., Ling, C., Zaki, M. J., Siebes, A., Yu, J. X., Goethals, B., Webb, G. I., & Wu, X. (Eds.), *ICDM Workshops*, pp. 687–692. IEEE Computer Society.

Lassen, D. S., & Brown, A. R. (2011). Twitter the electoral connection?. *Social Science Computer Review*, *29*(4), 419–436.

Lazaridou, A., Bruni, E., & Baroni, M. (2014). Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL*, pp. 1403–1414.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Li, J., Zhou, G., Wang, H., & Zhu, Q. (2010). Learning the scope of negation via shallow semantic parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pp. 671–679, Beijing, China.

Li, X. G., Li, S. M., Jiang, L. R., & Zhang, S. B. (2014). Study of english pronunciation quality evaluation system with tone and emotion analysis capabilities. *Applied Mechanics and Materials*, *475*, 318–323.

Lin, C., He, Y., & Everson, R. (2011). Sentence subjectivity detection with weakly-supervised learning.. In *IJCNLP*, pp. 1153–1161.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, *35*(03), 121–143.

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

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Aggarwal, C. C., & Zhai, C. (Eds.), *Mining Text Data*, pp. 415–463. Springer US.

Liu, H., Lieberman, H., & Selker, T. (2003a). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pp. 125–132, New York, NY. ACM.

Liu, H., Selker, T., & Lieberman, H. (2003b). Visualizing the affective structure of a text document. In *CHI'03 extended abstracts on Human factors in computing systems*, pp. 740–741. ACM.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). Low-quality product review detection in opinion summarization.. In *EMNLP-CoNLL*, pp. 334–342.

Lloret, E., Balahur, A., Palomar, M., & Montoyo, A. (2009). Towards building a competitive opinion summarization system: challenges and keys. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pp. 72–77. Association for Computational Linguistics.

Louviere, J. J. (1991). Best-worst scaling: A model for the largest difference judgments. Working Paper.

Lu, B., & Tsou, B. K. (2010). Cityu-dac: Disambiguating sentiment-ambiguous adjectives within context. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 292–295. Association for Computational Linguistics.

Ma, C., Prendinger, H., & Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In Tao, J., & Picard, R. W. (Eds.), *First International Conference on Affective Computing and Intelligent Interaction (ACII-2005)*, pp. 622–628, Beijing, China.

Makki, R., Brooks, S., & Milios, E. E. (2014). Context-specific sentiment lexicon expansion via minimal user interaction. In *Proceedings of the International Conference on Information Visualization Theory and Applications*, pp. 178–186, Rome, Italy.

Malti, T., & Krettenauer, T. (2013). The relation of moral emotion attributions to prosocial and antisocial behavior: A meta-analysis. *Child Development*, *84*(2), 397–412.

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureñalópez, L. A., & Montejoráez, A. R. (2012). Sentiment analysis in Twitter. *Natural Language Engineering*, 1–28.

Maynard, D., & Funk, A. (2011). Automatic detection of political opinions in tweets. *gateacuk*, *7117*, 81–92.

McNair, D., Lorr, M., & Droppleman, L. (1989). Profile of mood states (poms)..

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

Mihalcea, R., & Liu, H. (2006). A corpus-based approach to finding happiness. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 139–144. AAAI Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Minamikawa, A., & Yokoyama, H. (2011). Personality estimation based on weblog text classification. In *Modern Approaches in Applied Intelligence*, pp. 89–97. Springer.

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language*

Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 105–114, Portland, OR, USA. Association for Computational Linguistics.

Mohammad, S. (2012). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 587–591, Montréal, Canada.

Mohammad, S., Kiritchenko, S., & Zhu, X. (2013a). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA.

Mohammad, S., Kiritchenko, S., & Zhu, X. (2013b). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Mohammad, S., & Yang, T. (2011a). Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pp. 70–79, Portland, Oregon. Association for Computational Linguistics.

Mohammad, S., & Yang, T. (2011b). Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pp. 70–79, Portland, Oregon.

Mohammad, S. M. (2012a). #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pp. 246–255, Stroudsburg, PA.

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

Mohammad, S. M., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP '09, pp. 599–608.

Mohammad, S. M., & Kiritchenko, S. (2013). Using nuances of emotion to identify personality. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-13)*, Boston, MA.

Mohammad, S. M., & Kiritchenko, S. (2014). Using hashtags to capture fine emotion categories from tweets. *To appear in Computational Intelligence*.

Mohammad, S. M., Salameh, M., & Kiritchenko, S. (2015). How translation alters sentiment. In *Journal of Artificial Intelligence Research*.

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2014). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*.

Montero, C. S., Munezero, M., & Kakkonen, T. (2014). Investigating the role of emotion-based features in author gender classification of text. In *Computational Linguistics and Intelligent Text Processing*, pp. 98–114. Springer.

Murakami, A., & Raymond, R. (2010). Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Coling 2010: Posters*, pp. 869–875, Beijing, China. Coling 2010 Organizing Committee.

Nalisnick, E. T., & Baird, H. S. (2013a). Character-to-character sentiment analysis in shakespeares plays..

Nalisnick, E. T., & Baird, H. S. (2013b). Extracting sentiment networks from shakespeare's plays. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pp. 758–762. IEEE.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, pp. 278–281, San Jose, California.

Nielsen, F. Å. (2011). Afinn..

Orme, B. (2009). Maxdiff analysis: Simple counting, individual-level logit, and HB. Sawtooth Software, Inc.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The measurement of meaning*. University of Illinois Press.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, LREC '10, Valletta, Malta.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Parrot, W. (2001). *Emotions in Social Psychology*. Psychology Press.

Paul, M. J., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health.. In *ICWSM*, pp. 265–272.

Pearl, L., & Steyvers, M. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

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

Picard, R. W. (1997, 2000). *Affective computing*. MIT press.

Pit, M. (2006). Determining subjectivity in text: The case of backward causal connectives in dutch. *Discourse Processes*, *41*(2), 151–174.

Plutchik, R. (1962). *The Emotions*. New York: Random House.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, *1*(3), 3–33.

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

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

Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. In *Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 339–346, Stroudsburg, PA, USA.

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pp. 482–491, Stroudsburg, PA.

Qadir, A. (2009). Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Proceedings of the Workshop on Events in Emerging Text Types*, eETTs '09, pp. 38–43, Stroudsburg, PA, USA.

Quan, C., & Ren, F. (2009). Construction of a blog emotion corpus for chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pp. 1446–1454. Association for Computational Linguistics.

Quan, C., & Ren, F. (2014). Visualizing emotions from chinese blogs by textual emotion analysis and recognition techniques. *International Journal of Information Technology & Decision Making*, 1–20.

Ren, F., & Quan, C. (2012). Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, *13*(4), 321–332.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, *47*(1), 239–268.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105–112. Association for Computational Linguistics.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*.

Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Nakov, P., & Zesch, T. (Eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval-2014, Dublin, Ireland.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133.

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

Salameh, M., Mohammad, S. M., & Kiritchenko, S. (2015). Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the North American Chapter of Association of Computational Linguistics*, Denver, Colorado.

Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Lucas, R., Agrawal, M., Park, G., et al. (2013). Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach.. *Journal of personality and social psychology*, *52*(6), 1061.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, Seattle, USA.

Somasundaran, S., & Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pp. 226–234, Stroudsburg, PA, USA. Association for Computational Linguistics.

Somprasertsri, G., & Lalitrojwong, P. (2010). Mining feature-opinion in online customer reviews for opinion summarization.. *J. UCS*, *16*(6), 938–955.

Sridhar, D., Getoor, L., & Walker, M. (2014). Collective stance classification of posts in online debate forums. *ACL 2014*, 109.

Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., & associates (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Stoyanov, V., & Cardie, C. (2006). Toward opinion summarization: Linking the sources. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pp. 9–14. Association for Computational Linguistics.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pp. 70–74, Prague, Czech Republic.

Su, F., & Markert, K. (2008). From words to senses: a case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 825–832. Association for Computational Linguistics.

Su, Q., Xiang, K., Wang, H., Sun, B., & Yu, S. (2006). Using pointwise mutual information to identify implicit features in customer reviews. In *Proceedings of the 21st international conference on Computer Processing of Oriental Languages: beyond the orient: the research challenges ahead*, ICCPOL'06, pp. 22–30, Berlin, Heidelberg. Springer-Verlag.

Suero Montero, C., & Suhonen, J. (2014). Emotion analysis meets learning analytics: online learner profiling beyond numerical data. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research*, pp. 165–169. ACM.

Sun, Y., Quan, C., Kang, X., Zhang, Z., & Ren, F. (2014). Customer emotion detection by emotion expression analysis on adverbs. *Information Technology and Management*, 1–9.

Suttles, J., & Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pp. 121–136. Springer.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Taboada, M., Voll, K., & Brooke, J. (2008). Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

Tang, D., Wei, F., Qin, B., Liu, T., & Zhou, M. (2014a). Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 208–212.

Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2014b). Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of COLING*, pp. 172–182.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Thomas, B., et al. (2014). Synthesized feature space for multiclass emotion classification. In *Networks & Soft Computing (ICNSC), 2014 First International Conference on*, pp. 188–192. IEEE.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 327–335, Sydney, Australia.

Tokuhisa, R., Inui, K., & Matsumoto, Y. (2008). Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pp. 881–888.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010a). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*(4), 402–418.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010b). Predicting elections with Twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, 178–185.

Turney, P., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, *21*(4).

Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons.. In *ECAI*, Vol. 215, pp. 765–770.

Vo, B., & Collier, N. (2013). Twitter emotion analysis in earthquake situations. *International Journal of Computational Linguistics and Applications*, *4*(1), 159–173.

Walker, M. A., Anand, P., Abbott, R., & Grant, R. (2012). Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 592–596.

Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 553–561.

Wang, C., & Wang, F. (2012). A bootstrapping method for extracting sentiment words using degree adverb patterns. In *Computer Science & Service System (CSSS), 2012 International Conference on*, pp. 2173–2176. IEEE.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pp. 587–592, Washington, DC, USA. IEEE Computer Society.

Wang, X., & Fu, G.-H. (2010). Chinese subjectivity detection using a sentiment density-based naive bayesian classifier. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, Vol. 6, pp. 3299–3304. IEEE.

Wang, Z. (2014). Segment-based fine-grained emotion detection for chinese text. *CLP 2014*, 52.

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

Wen, S., & Wan, X. (2014). Emotion classification in microblog texts using class sequential rules. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, pp. 486–497. Springer.

Wiebe, J., & Riloff, E. (2011). Finding mutual benefit between subjectivity analysis and information extraction. *Affective Computing, IEEE Transactions on*, *2*(4), 175–191.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational linguistics*, *30*(3), 277–308.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pp. 60–68, Stroudsburg, PA, USA.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pp. 34–35. Association for Computational Linguistics.

Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., & Ritter, A. (2013). SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354, Stroudsburg, PA, USA.

Xu, G., Huang, C.-R., & Wang, H. (2013). Extracting chinese product features: representing a sequence by a set of skip-bigrams. In *Proceedings of the 13th Chinese conference on Chinese Lexical Semantics*, CLSW'12, pp. 72–83, Berlin, Heidelberg. Springer-Verlag.

Xu, R., Wong, K.-F., Lu, Q., Xia, Y., & Li, W. (2008). Learning knowledge from relevant webpage for opinion analysis. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, Vol. 1, pp. 307–313. IEEE.

Yen, H. Y., Lin, P. H., & Lin, R. (2014). Emotional product design and perceived brand emotion..

Yu, L.-C., Wu, J.-L., Chang, P.-C., & Chu, H.-S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*, 89–97.

Zhang, C., Zeng, D., Xu, Q., Xin, X., Mao, W., & Wang, F.-Y. (2008). Polarity classification of public health opinions in chinese. In *Intelligence and Security Informatics*, pp. 449–454. Springer.

Zhang, L., Liu, B., Lim, S. H., & O'Brien-Strain, E. (2010). Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pp. 1462–1470, Stroudsburg, PA, USA.

Zhe, X., & Boucouvalas, A. (2002). *Text-to-Emotion Engine for Real Time Internet CommunicationText-to-Emotion Engine for Real Time Internet Communication*, pp. 164–168.

Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014a). An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 304–313, Baltimore, Maryland.

Zhu, X., Kiritchenko, S., & Mohammad, S. M. (2014b). NRC-Canada-2014: Recent improvements in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Zhu, X., Sobhani, P., & Guo, H. (2015). Long short-term memory over recursive structures. In *Proceedings of International Conference on Machine Learning*.