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

Abstract

Recent advances in machine learning have led to computer systems that are human-like in behaviour. Sentiment analysis, the automatic determination of emotions in text, is allowing us to capitalize on substantial previously unattainable opportunities in commerce, public health, government policy, social sciences, and art. Further, analysis of emotions in text, from news to social media posts, is improving our understanding of not just how people convey emotions through language but also how emotions shape our behaviour. This article presents a sweeping overview of sentiment analysis research that includes: the origins of the field, the rich landscape of tasks, challenges, a survey of the methods and resources used, and applications. We also discuss discuss how, without careful fore-thought, sentiment analysis has the potential for harmful outcomes. We outline the latest lines of research in pursuit of fairness in sentiment analysis.

Keywords: *sentiment analysis, emotions, artificial intelligence, machine learning, natural language processing (NLP), social media, emotion lexicons, fairness in NLP*

1. Introduction

Sentiment analysis is an umbrella term for the determination of valence, emotions, and other affectual states from text or speech automatically using computer algorithms. Most commonly, it is used to refer to the task of automatically determining the valence of a piece of text, whether it is positive, negative, or neutral, the star rating of a product or movie review, or a real-valued score in a 0 to 1 range that indicates the degree of positivity of a piece of text. However, more generally, it can refer 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. Sentiment analysis can also refer to the task of determining one's emotional state from their utterances (irrespective of whether the text is expressing an attitude towards an outside entity). The name *sentiment analysis* is a legacy of early work that focused heavily on appraisals in customer reviews. Since its growth to encompass emotions and feelings, some now refer to the field more broadly as *emotion analysis*.

Sentiment analysis, as a field of research, arose at the turn of the century with the publication of some highly influential Natural Language Processing (NLP) research.[1] This

---

1. See this series of blog posts for a visual analysis of NLP papers, conferences, and journals from 1965 to 2019: https://medium.com/@nlpscholar/state-of-nlp-cbf768492f90. Note that conference papers in NLP (as in Computer Science) are fully peer-reviewed and published in the proceedings. They are also the predominant mode of publication; far outnumbering journal publications. The top NLP conferences are markedly competitive, with roughly one in four or five submissions being accepted.

initial work was on determining the valence (or sentiment) in customer reviews (Turney, 2002; Pang, Lee, & Vaithyanathan, 2002) and on separating affectual or subjective text from more factual and non-affective text (Wiebe, Bruce, & O'Hara, 1999; Riloff & Wiebe, 2003a). The platform for this work was set in the 1990's with the growing availability of large amounts of digitally accessible text as well as, of course, earlier seminal work at the intersection of emotions and language in psychology, psycolinguistics, cognitive psychology, and behavioural science (Osgood, Suci, & Tannenbaum, 1957; Russell, 1980; Oatley & Johnson-Laird, 1987). See Figure 1 (a) for a timeline graph of the number of papers in the ACL Anthology with sentiment and associated terms in the title. (The ACL Anthology is a repository of public domain, free to access, articles on NLP with more than 50K articles published since 1965.[2]) Figure 1 (b) shows a timeline graph where each paper is represented as a segment whose height is proportional to the number of citations it has received as of June 2019.[3] Observe the large citations impact of the 2002 papers by Turney (2002) and Pang et al. (2002) and subsequent papers published in the early to mid 2000s. Since then, the number of sentiment analysis papers published every year has steadily increased.

Several notable works on the relationship between language and sentiment were carried out much earlier, though. Osgood et al. (1957) asked human participants to rate words along dimensions of opposites such as *heavy–light, good–bad, strong–weak,* etc. Their seminal work on these judgments showed that the most prominent dimension of connotative word meaning is evaluation (aka valence) (*good–bad, happy–sad*), followed by potency (*strong–weak, dominant–submissive*) and activity (*active–passive, energetic–sluggish*). Russell (1980) showed through similar analyses of emotion words that the three primary dimensions of emotions are valence (pleasure–displeasure, happy–sad), arousal (active–passive), and dominance (dominant–submissive). Ortony, Clore, and Collins (1988) argued 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). Even though Osgood, Russell, and Ortony chose slightly different words as representative of the primary dimensions, e.g. evaluation vs. valence, potency vs. dominance, and activity vs. arousal, their experiments essentially led to the same primary dimensions of emotion or connotative meaning. For the purposes of this chapter, we will use the terms coined by Russell (1980): valence (V), arousal (A), and dominance (D).

In contrast to the VAD model, Paul Ekman and others developed theories on how some emotions are more basic, more important, or more universal, than others (Ekman, 1992; Ekman & Friesen, 2003; Plutchik, 1980, 1991). These emotions, such as joy, sadness, fear, anger, etc., are argued to have ties to universal facial expressions and physiological processes such as increased heart rate and perspiration. Ekman (1992), Plutchik (1980), Parrot (2001), and Frijda (1988) and others proposed different sets of basic emotions. However, these assertions of universal ties between emotions and facial expressions, as well as the

---

2. https://www.aclweb.org/anthology/
3. Terms used for the query in the visualization: *sentiment, valence, emotion, emotions, affect, polarity, subjective, subjectivity,* and *stance.* NLP Scholar: http://saifmohammad.com/WebPages/nlpscholar.html

A1. #papers

A2. #papers, by year of publication

B1. #citations

B2. #citations, by year of publication

C. Papers

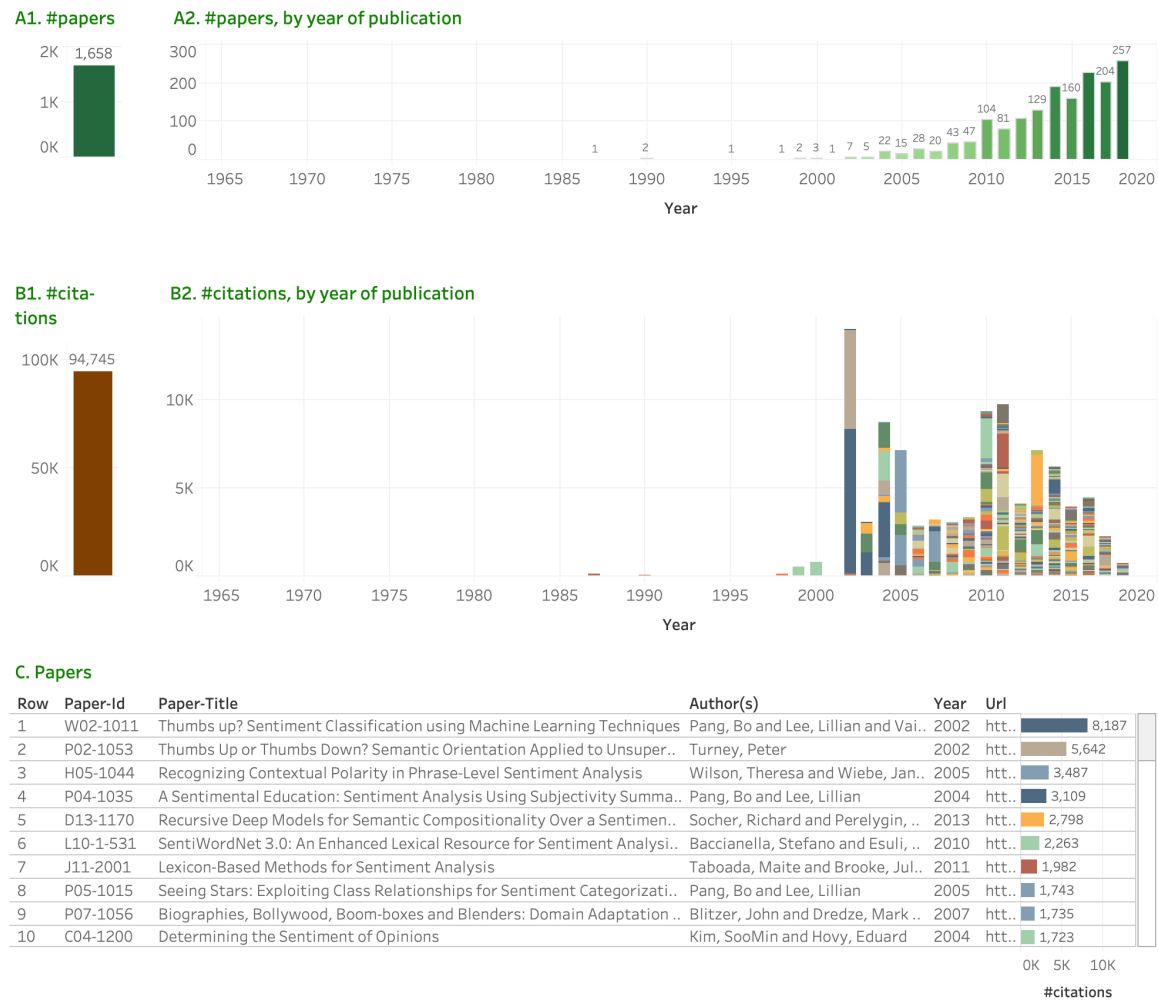| Row	Paper-Id	Paper-Title	Author(s)	Year	Url	#citations
1	W02-1011	Thumbs up? Sentiment Classification using Machine Learning Techniques	Pang, Bo and Lee, Lillian and Vai..	2002	htt..	8,187
2	P02-1053	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsuper..	Turney, Peter	2002	htt..	5,642
3	H05-1044	Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis	Wilson, Theresa and Wiebe, Jan..	2005	htt..	3,487
4	P04-1035	A Sentimental Education: Sentiment Analysis Using Subjectivity Summa..	Pang, Bo and Lee, Lillian	2004	htt..	3,109
5	D13-1170	Recursive Deep Models for Semantic Compositionality Over a Sentimen..	Socher, Richard and Perelygin, ..	2013	htt..	2,798
6	L10-1-531	SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysi..	Baccianella, Stefano and Esuli, ..	2010	htt..	2,263
7	J11-2001	Lexicon-Based Methods for Sentiment Analysis	Taboada, Maite and Brooke, Jul..	2011	htt..	1,982
8	P05-1015	Seeing Stars: Exploiting Class Relationships for Sentiment Categorizati..	Pang, Bo and Lee, Lillian	2005	htt..	1,743
9	P07-1056	Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation ..	Blitzer, John and Dredze, Mark ..	2007	htt..	1,735
10	C04-1200	Determining the Sentiment of Opinions	Kim, SooMin and Hovy, Eduard	2004	htt..	1,723

Figure 1: A screenshot of the NLP Scholar dashboard when the search box is used to show only those papers that have a sentiment analysis associated term in the title. A1 shows the number of number of papers, and A2 shows the number of papers by year of publication. B1 shows the number of citations for the set (as of June 2019) and B2 shows the citations by year of publication. For a given year, the bar is partitioned into segments corresponding to individual papers. Each segment (paper) has a height that is proportional to the number of citations it has received and assigned a colour at random. C shows the list of papers ordered by number of citations. Hovering over a paper in the interactive visualization shows metadata associated with the paper.

theory of basic emotions have been challenged and hotly debated in recent work (Barrett, 2006; De Leersnyder, Boiger, & Mesquita, 2015).

This chapter presents a comprehensive overview of work on automatically detecting valence, emotions, and other affectual states from *text*.[4] We begin in Section 2 by discussing various challenges to sentiment analysis, including: the subtlety with emotions can be conveyed, the creative use of language, the bottleneck of requiring annotated data, and the lack of para-linguistic context. 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.

Sections 4 delves into approaches to automatically detect emotions from text (especially, sentences and tweets). We discuss the broad trends in machine learning that have swept across Natural Language Processing (NLP) and sentiment analysis, including: transfer learning, deep neural networks, and the representation of words and sentences with dense, low-dimensional, vectors. We also identify influential past and present work, along with influential annotated resources for affect prediction. Section 5 summarizes work on creating emotion lexicons, both manually and automatically. Notably, we list several influential lexicons pertaining to the basic emotions as well as valence, arousal, and dominance. Section 6 explores work on determining the impact of sentiment modifiers, such as negations, intensifiers, and modals.

Section 7 discusses some preliminary sentiment analysis work on sarcasm, metaphor, and other 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.

Section 9 presents prominent areas where sentiment analysis is being applied: from commerce and intelligence gathering to policy making, public, health, and even art. However, it should be noted that these applications have not always yielded beneficial results. Recent advances in machine learning have meant that computer systems are becoming more human-like in their behaviour. This also means that they perpetuate human biases. Some learned biases may be beneficial for the downstream application. Other biases can be inappropriate and result in negative experiences for some users. Examples include, loan eligibility and crime recidivism systems that negatively assess people belonging to a certain area code (which may disproportionately impact people of certain races) (Chouldechova, 2017) and resumé sorting systems that believe that men are more qualified to be programmers than women (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). Similarly, sentiment analysis systems can also perpetuate and accentuate inappropriate human biases. We discuss issues of fairness and bias in sentiment analysis in Section 10.

---

4. 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 introduction of giving machines the ability to detect sentiment and emotions in various modalities such as text, speech, and vision. See articles by Lawrence and Reed (2020) and Cabrio and Villata (2018) for surveys on the related area of argument mining.

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 stated explicitly. For example:

    Another Monday, and another week working my tail off.

    conveys a sense of frustration without the speaker explicitly saying so. Note also 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, including 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 terms commonly seen in social media.

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).

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. The degree of agreement between annotators is significantly lower when 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 world wide web to hundreds or thousands of unknown respondents. Section 5.1.2 describes an annotation scheme called *best worst scaling (BWS)* that has led to more high-quality and consistent sentiment annotations.[5]

In the sections ahead we describe approaches that, to some extent, address these issues. Nonetheless, significant progress still remains to be made.

---

5. Maximimum Difference Scaling (max-diff) is a slight variant of BWS, and sometimes the terms are used interchangeably.

3. Sentiment Analysis Tasks

Even though early work focused extensively on the sentiment in customer reviews, sentiment analysis involves a diverse landscape of tasks. Thus, when building or using sentiment analysis systems, it is important to first determine the precise task that is intended: for example, whether it is to determine the sentiment of the reader or the writer, whether it is to determine attitude towards a product or one's emotional state, whether it is intended to assess sentences, tweets, or documents, etc. We discuss the landscape of sentiment analysis tasks below.

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 that 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 overt 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 of text: from sentiment associations of words and phrases; to sentiments of sentences, SMS messages, chat messages, and tweets. One can also explore sentiment in product reviews, blog posts, whole documents, and even streams of texts such as tweets mentioning an entity over time.

Words: Some words denotatate valence, i.e., valence is part of their core meaning, for example, *good, bad, terrible, excellent, nice,* and so on. Some other words do not denotate

valence, but have strong positive or negative associations (or connotations). For example, *party* and *raise* are associated with positive valence, whereas *slave* and *death* are associated with negative valence. Words that are not strongly associated with positive or negative 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 have 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 (Duppada, Jain, & Hiray, 2018; Agrawal & Suri, 2019; Huang, Trabelsi, & Zaïane, 2019; Abdou, Kulmizev, & Ginés i Ametllé, 2018; Mohammad, Kiritchenko, & Zhu, 2013a; Kiritchenko, Zhu, Cherry, & Mohammad, 2014a; Zhu, Kiritchenko, & Mohammad, 2014b; Tang, Wei, Qin, Liu, & Zhou, 2014a) in recent sentiment-related shared tasks, The SemEval-2018 Affect in Tweets, SemEval-2013 and

2014 Sentiment Analysis in Twitter, used large sentiment lexicons (Mohammad, Bravo-Marquez, Salameh, & Kiritchenko, 2018; Chatterjee, Narahari, Joshi, & Agrawal, 2019; Wilson, Kozareva, Nakov, Rosenthal, Stoyanov, & Ritter, 2013; Rosenthal, Nakov, Ritter, & Stoyanov, 2014).[6] (Some of these tasks also had separate 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 and emotion 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) as well as a large amount of (unlabeled) raw text using low-dimensional vector representations of the sentences and constituent words, as well as more traditional features such as those drawn from word and character ngrams, valence and emotion association lexicons, and negation lists. We discuss the vaelnce (sentiment) and emotion classification systems and the resources they commonly use in Sections 4.2 and 4.3, respectively.

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; Somprasertsri & Lalitrojwong, 2010) and on analyzing patterns of sentiment in social networks in novels and fairy tales (Nalisnick & Baird, 2013; Mohammad & Yang, 2011a; Davis & Mohammad, 2014a).

Document and Twitter Streams: Sentiment analysis has also been applied to streams of documents and twitter streams where the purpose is usually to detect aggregate trends in emotions over time. This includes work on determining media portrayal of events by analyzing online news streams (Liu, Gulla, & Zhang, 2016) and predicting stock market trends through the sentiment in financial news articles (Schumaker, Zhang, Huang, & Chen, 2012), finance-related blogs (O'Hare, Davy, Bermingham, Ferguson, Sheridan, Gurrin, & Smeaton, 2009), and tweets (Smailović, Grčar, Lavrač, & Žnidaršič, 2014). There is also considerable interest in tracking public opinion over time: Thelwall, Buckley, and Paltoglou (2011a) tracked sentiment towards 30 events in 2010 including oscars, earth quake, tsunami, and celebrity scandals; Bifet, Holmes, Pfahringer, and Gavalda (2011) analyzed sentiment in tweets pertaining to the 2010 Toyota crisis; Kim, Jeong, Kim, Kang, and Song (2016) tracked sentiment in Ebola-related tweets; Vosoughi, Roy, and Aral (2018) published influential work that showed that false stories elicited responses of fear, disgust, and surprise, whereas true stories elicited responses with anticipation, sadness, joy, and trust; Fraser, Zeller, Smith, Mohammad, and Rudzicz (2019) analyzed emotions in the tweets mentioning a hitch-hiking Canadian robot; and it is inevitable that soon there will be published work on tracking various aspects of the global Covid-19 pandemic.

It should be noted that the analyses of streams of data have unique challenges, most notable in the drift of topic of discussion. Thus for example, if one is to track the tweets pertaining to the Covid-19 pandemic, one would have started by polling the Twitter API for tweets with hashtags #coronavirus and #wuhanvirus, but will have to update the the search

---

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

hashtags continually over time to terms such as #covid19, #socialdistancing, #physicald-istancin, #UKlockdown, and so on. Yet another, notable challenge in aggregate analysis such as that in determining public emotional state through social media are the biases that impact positing behaviour. For example, it is well known that people have a tendency to talk more about positive feeling and experiences to show that they are happy and successful (Meshi, Morawetz, & Heekeren, 2013; Jordan, Monin, Dweck, Lovett, John, & Gross, 2011). In contrast, when reporting about products there is a bias towards reporting shortcomings and negative experiences (Hu, Zhang, & Pavlou, 2009). It is also argued that there is a bias in favor of report extreme high-arousal emotions because such posts are more likely to spread and go viral (Berger, 2013).

An offshoot of work on the sentiment analysis of streams of text is work on visualizing emotion in these text streams Lu, Hu, Wang, Kumar, Liu, and Maciejewski (2015) visualized sentiment in geolocated ebola tweets; Gregory, Chinchor, Whitney, Carter, Hetzler, and Turner (2006) visualized reviews of *The Da Vinci Code*; and Fraser et al. (2019) visualized emotions in tweets before and after the death of HitchBot and across countries. See survey articles by Boumaiza (2015) and Kucher, Paradis, and Kerren (2018) for further information on sentiment visualization techniques.

3.3 Detecting Sentiment Towards a Target

3.3.1 Detecting Sentiment Towards Aspects of an Entity or Aspect Based Sentiment Analysis (ABSA)

A review of a product or service can express sentiment towards various aspects. For example, a restaurant review can praise the food served, but express anger towards the quality of service. There is a large body of work on detecting aspects of products and also sentiment towards these aspects (Schouten & Frasincar, 2015; 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 and 2016. It will be useful to develop aspect-based sentiment systems for other domains such as blogs and news articles as well. (See proceedings of the ABSA tasks in SemEval-2014, 2015, and 2016 for details about participating aspect sentiment systems.[7]) As with various task in NLP, over the last year, a number of distant learning based approaches for ABSA have been proposed, for example Li, Bing, Zhang, and Lam (2019) develop a BERT-based benchmark for the task. See surveys by Do, Prasad, Maag, and Alsadoon (2019), Laskari and Sanampudi (2016), Schouten and Frasincar (2015), Vohra and Teraiya (2013) for further information on aspect based sentiment analysis.

---

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

3.3.2 DETECTING STANCE

Stance detection is the task of automatically determining from text whether the author of the text is in favor or against a proposition or target. Early work in stance detection, focused 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). However, more recently, there has been a spurt of work on social media texts. The first shared task on detecting stance in tweets was organized in 2016 Mohammad, Kiritchenko, Sobhani, Zhu, and Cherry (2016), Mohammad, Sobhani, and Kiritchenko (2017) . They framed the task as follows: Given a tweet text and a pre-determined target proposition, state whether the tweeter is likely in favor of the proposition, against the proposition, or whether neither inference is likely. For example, given the following target and text pair:

> Target: *Pro-choice movement OR women have the right to abortion*
> Text: *A foetus has rights too!*

Automatic systems have to deduce the likely stance of the tweeter towards the target. 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 existing unlabeled text about the target.[8]

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. Mohammad et al. (2017) analyzed manual annotations for stance and sentiment on the same data for a number of targets.

Even though about twenty teams participated in the first shared task on stance, including some that used the latest recursive neural network models, none could do better than a simple support vector machine baseline system put up by the organizers that used word and character n-gram features (Sobhani, Mohammad, & Kiritchenko, 2016). Nonetheless, there has been considerable followup work since then (Yan, Chen, & Shyu, 2020; Xu, Mohtarami, & Glass, 2019; Darwish, Stefanov, Aupetit, & Nakov, 2019), including new stance shared

---

8. Here 'unlabeled' refers to text that is not labeled for stance.

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.

tasks for Catalan (Taulé, Martí, Rangel, Rosso, Bosco, Patti, et al., 2017) and work on jointly modeling stance and sentiment (Ebrahimi, Dou, & Lowd, 2016). See survey article by Kücük and Can (2020) for more details.

Automatically detecting stance has widespread applications in information retrieval, text summarization, and textual entailment. One can also argue that stance detection is more useful in commerce, brand management, public health, and policy making than simply identifying whether the language used in a piece of text is positive or negative.

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.[9] 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

---

9. https://framenet.icsi.berkeley.edu/fndrupal/home

available for download.[10] 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 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), newspaper headlines (Bellegarda, 2010), novels (Boucouvalas, 2002; John, Boucouvalas, & Xu, 2006; Francisco & Gervás, 2006; Mohammad & Yang, 2011a), emails (Liu, Lieberman, & Selker, 2003a; Mohammad & Yang, 2011a), blogs (Neviarouskaya, Prendinger, & Ishizuka, 2009; Genereux & Evans, 2006; Mihalcea & Liu, 2006), and tweets (Pak & Paroubek, 2010; Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Thelwall, Buckley, & Paltoglou, 2011b; 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.

Note also that the text genre often determines the kind of sentiment analysis suitable with it: for example, customer reviews are more suited to determine how one feels about a product than to determine one's general emotional state; personal diary type blog posts (and tweets) are useful for determining the writer's emotional state;[11] movie dialogues and fiction are well suited to identify the emotions of the characters (and not as much of the writer's emotional state); reactions to a piece of text (for example, the amount of likes, applause, or other reactions) are useful in studying the reader's attitude towards the text; and so on.

4.1 Machine Learning Algorithms

Given some text and associated true emotion labels (commonly referred to as *training data*), machine learning systems learn a model. (The emotion labels are usually obtained through annotations performed by native speakers of the language.) Then, given new previously unseen text, the model predicts its emotion label. The performance of the model is determined through its accuracy on a held out *test set* for which emotion labels are available as well.

For both the training and prediction phases, the sentences are first represented by a vector of numbers. These vectors can be a series of 0s and 1s or a series of real-valued

---

10. Political Tweets Dataset: www.purl.org/net/PoliticalTweets
11. Mohammad et al. (2015) showed that tweets predominantly convey the emotions of the tweeter.

numbers. Much of the work in the 1990s and 2000s represented sentences by carefully hand-engineered vectors, for example, whether the sentence has a particular word, whether the word is listed as a positive term in the sentiment lexicon, whether a positive word is preceded by a negation, the number of positive words in a sentence, and so on. The number of features can often be as large as hundreds of thousands. The features were mostly integer valued (often binary) and sparse, that is, a vector is composed mostly of zeroes and just a few non-zero integer values.

However, the dominant methodology since the 2010s is to represent words and sentences using vectors made up of only a few hundred real-valued numbers. These continuous word vectors, also called *embeddings*, are learned from deep neural networks (a type of machine learning algorithm). However, these approaches tend to require very large amounts of training data. Thus, recent approaches to deep learning rely heavily on another learning paradigm called *transfer learning*. The systems first learn word and sentence representations from massive amounts of raw text. For example, BERT (Devlin, Chang, Lee, & Toutanova, 2019)—one of the most popular current approaches, is trained on the entire Wikipedia corpus (about 2,500 million words) and a corpus of books (about 800 million words). Roughly speaking, the learning of the representations is driven by the idea that a good word or sentence representation is one that can be used to best predict the words or sentences surrounding it in text documents. Once these sentence representations are learned, they may be tweaked by a second round of learning that uses a small amount of task-specific training data; for example a few thousand emotion-labeled sentences.

Influential work on low-dimensional continuous representations of *words* includes models such as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), fastText (Bojanowski, Grave, Joulin, & Mikolov, 2016), and their variations (Collobert, Weston, Bottou, Karlen, Kavukcuoglu, & Kuksa, 2011; Le & Mikolov, 2014; Bojanowski, Grave, Joulin, & Mikolov, 2017; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018). Eisner, Rocktäschel, Augenstein, Bošnjak, and Riedel (2016) presented work on representing emoji with embeddings using a large corpora of tweets. ELMo (Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018) introduced a method to determine context-sensitive word representations, that is, instead of generating a fixed representation of words, it generates a different representation for every different context that the word is seen in.

Influential work on low-dimensional continuous representations of *sentences* includes models such as ULM-FiT (Howard & Ruder, 2018), BERT (Devlin et al., 2019), XLNet (Yang, Dai, Yang, Carbonell, Salakhutdinov, & Le, 2019), GPT-2 (Radford, Wu, Child, Luan, Amodei, & Sutskever, 2019) and their variations such as DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019), and RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, & Stoyanov, 2019).[12] Yet another variation, SentiBERT (Yin, Meng, & Chang, 2020), was developed to help in tasks such as determining the sentiment of larger text units from the sentiment of its constituents. A compilation of work on BERT and its variants is available online.[13]

---

12. Note that BERT and other related approaches can be used not only to generate sentence representations, but also to generate context-sensitive word representations (similar to ELMo).

13. https://github.com/tomohideshibata/BERT-related-papers/blob/master/README.md#multi-lingual

Despite the substantial dominance of neural and deep learning techniques in NLP research, more traditional machine learning frameworks, such as linear regression, support vector machines, and decision trees, as well as more traditional mechanisms to represent text features, such as word n-grams and sentiment lexicons, remain relevant. This is notably because of their simplicity, ease of use, interpretability, and even their potential ability to complement neural representations and improve prediction accuracy. 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 sentiment classification as well. Features from manually and automatically created sentiment lexicons such as word–valence association lexicons, word–arousal association lexicons, word–emotion association lexicons are also commonly used in conjunction with neural representations for detecting emotions or for detecting affect-related classes such as personality traits and states of well-being (Mohammad & Bravo-Marquez, 2017; Mohammad et al., 2018; Chatterjee et al., 2019; Agrawal & Suri, 2019). Examples of commonly used manually created sentiment lexicons are: the General Inquirer (GI) (Stone, Dunphy, Smith, Ogilvie, & associates, 1966), the NRC Emotion Lexicon (Mohammad & Turney, 2010; Mohammad & Yang, 2011a), the NRC Valence, Arousal, and Dominance Lexicon (Mohammad, 2018a), and VADER (Hutto & Gilbert, 2014). Commonly used automatically generated sentiment lexicons include SentiWordNet (SWN) (Esuli & Sebastiani, 2006), Sentiment 140 lexicon (Mohammad, Kiritchenko, & Zhu, 2013), and NRC Hashtag Sentiment Lexicon (Mohammad & Kiritchenko, 2014). Other traditional text features include those derived from parts of speech, punctuations (!, ???), word clusters, syntactic dependencies, negation terms (*no, not, never*), and word elongations (*hugggs, ah-hhh*). We will discuss manually and automatically generated emotion lexicons in more detail in Section 5.

4.2 Detecting Valence

There is tremendous interest in automatically determining valence in sentences and tweets through supervised machine learning systems. This is evident from the large number of research papers, textual datasets, shared task competitions, and machine learning systems developed for valence prediction over the last decade.

Shared Tasks

A number of shared task competitions on valence prediction have focused on tweets. These include the 2013, 2014, 2015, and 2016 SemEval shared tasks titled *Sentiment Analysis in Twitter (SAT)* and the 2018 Shared task titled *Affect in tweets*, the 2015 SemEval shared task *Sentiment Analysis of Figurative Language in Twitter*, and the SemEval-2020 shared task on Sentiment Analysis for Code-Mixed Social Media Text. Shared tasks on other sources of text include the SemEval-2019 Task 3: Contextual Emotion Detection in Text, the 2014, 2015, and 2016 SemEval shared tasks on *Aspect Based Sentiment Analysis (ABSA)*, and the 2015 Kaggle competition *Sentiment Analysis on Movie Reviews*.[14] Many of these

---

14. http://aclweb.org/aclwiki/index.php?title=SemEval_Portal
   SemEval-2015 Task 10: http://alt.qcri.org/semeval2015/task10/
   SemEval-2015 Task 11: http://alt.qcri.org/semeval2015/task11/
   SemEval-2015 Task 12: http://alt.qcri.org/semeval2015/task12/
   SemEval-2016 Task 4: http://alt.qcri.org/semeval2016/task4/

tasks received submissions from more than 40 teams from universities, research labs, and companies across the world. The SemEval-2018 Task 1: Affect in Tweets, is particularly notable, because it includes an array of subtasks on inferring both emotion classeses and emotion intensity, provides labeled data for English, Arabic, and Spanish tweets, and for the first time in an NLP shared task, analyzed systems for bias towards race and gender (Mohammad et al., 2018). (We discuss issues pertaining to ethics and fairness further in the last section of this chapter.)

Systems

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. Notably, 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.

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), Zhu, Sobhani, and Guo (2015), Wang, Yu, Lai, and Zhang (2016), and others further explored the use recursive neural networks and word embeddings in sentiment analysis. Recent work on Aspect Based Sentiment Analysis has also explored using recurrent neural network (RNNs) and long short-term memory (LSTM) (particularly types of neural network) (Tang, Qin, Feng, & Liu, 2016; Chen, Sun, Bing, & Yang, 2017; Ruder, Ghaffari, & Breslin, 2016).

More recent works have explored transfer learning for sentiment analysis (Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019). Systems built for shared task competitions (Duppada et al., 2018; Agrawal & Suri, 2019; Huang et al., 2019; Abdou et al., 2018) often make use of ensemble neural systems, most commonly building on pre-tarined BERT, ULMFit, GloVe, and word2vec, followed by fine-tuning the system on the provided labeled training set.

Even though the vast majority of sentiment analysis work is on English datasets, there is growing research in Chinese (Zhou, Lu, Dai, Wang, & Xiao, 2019; Xiao, Li, Wang, Yang,

---

SemEval-2016 Task 5: http://alt.qcri.org/semeval2016/task5/

SemEval-2018 Task 1: https://competitions.codalab.org/competitions/17751

SemEval-2020 Task 3: https://competitions.codalab.org/competitions/19790

SemEval-2020 Task 9: https://competitions.codalab.org/competitions/20654

http://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

Du, & Sangaiah, 2018; Wan, 2008) and Arabic dialects (Dahou, Xiong, Zhou, Haddoud, & Duan, 2016; El-Naggar, El-Sonbaty, & El-Nasr, 2017; Gamal, Alfonse, El-Horbaty, & Salem, 2019). (Further details on Chinese sentiment analysis can be found in this survey article by Peng, Cambria, and Hussain (2017); Further details on Arabic sentiment analysis can be found in these survey articles by Al-Ayyoub, Khamaiseh, Jararweh, and Al-Kabi (2019), Alowisheq, Alhumoud, Altwairesh, and Albuhairi (2016).) There is some work on other European and Asian languages, but little or none on native African and indigenous languages from around the world.

4.3 Automatically Detecting and Analyzing Emotions

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. Asking annotators to label the data 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. 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 and Plutchik's Eight**: Paul Ekman proposed a set of six basic emotions that include: joy, sadness, anger, fear, disgust, and surprise (Ekman, 1992; Ekman & Friesen, 2003). Robert Plutchik's set of basic emotions includes Ekman's six as well as trust and anticipation. Figure 2 shows how Plutchik arranges 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) for details about his taxonomy of emotions created by primary, secondary, and tertiary emotions.

  Since the writing style and vocabulary in different sources, such as chat messages, blog posts, and newspaper articles, can be very different, automatic systems that cater to specific domains are more accurate when trained on data from the target domain. Alm, Roth, and Sproat (2005) annotated 22 Grimm fairy tales (1580 sentences) for Ekman emotions.[15] Strapparava and Mihalcea (2007) annotated newspaper headlines with intensity scores for each of the Ekman emotions, referred to as the *Text Affect Dataset*.[16] Aman and Szpakowicz (2007) annotated blog posts with the Ekman emotions. Mohammad (2012) experimented on the Text Affect and Aman datasets to show that word ngram features tend to be data and domain specific, and thus not useful when working on a test set from a different domain, whereas, emotion lexicon features are more robust and apply more widely.

  Mohammad (2012a) polled the Twitter API for tweets that have hashtag words such as *#anger* and *#sadness* corresponding to the eight Plutchik emotions.[17] He showed

---

15. https://dl.dropboxusercontent.com/u/585585/RESOURCEWEBSITE1/index.html
16. http://web.eecs.umich.edu/~mihalcea/downloads.html#affective
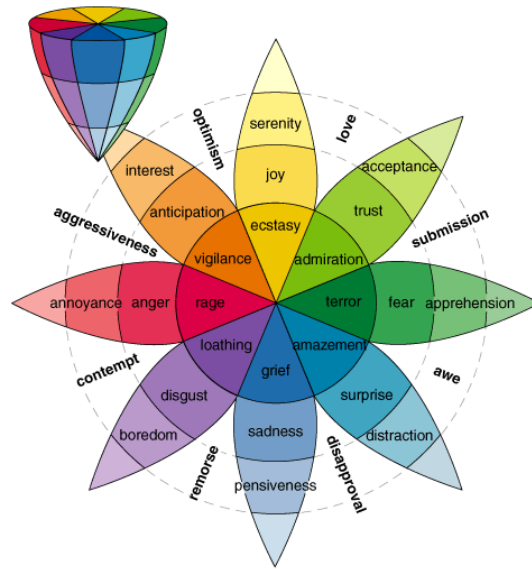17. http://saifmohammad.com/WebPages/lexicons.html

Figure 2: Set of eight basic emotions proposed in Plutchik (1980).

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.

As part of the SemEval-2018 Task 1: Affect in Tweets, Mohammad et al. (2018) created labeled datasets of English, Arabic, and Spanish tweets. The labels include: emotion intensity scores, emotion intensity ordinal classification, and emotion classes, in addition to valence annotations. As part of the SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text, Chatterjee et al. (2019) provide a dataset of dialogues between a user and computer agents that are annotated for the emotions of the user.

- **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.[18] 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.[19] Experience Project is a portal where users share their life experiences. These shared texts are usually rife with emotion.

Bollen, Pepe, and Mao (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.

- **Distant Supervision Datasets:** Distant supervision techniques proposed in Mohammad (2012a), Purver and Battersby (2012), Wang, Chen, Thirunarayan, and Sheth (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. Pool and Nissim (2016) collected public Facebook posts and their associated user-marked emotion reaction data from news and entertainment companies to develop an emotion classification system. Felbo, Mislove, Søgaard, Rahwan, and Lehmann (2017) collected millions of tweets with emojis and developed neural transfer learning models from them that were applied to sentiment and emotion detection tasks. Abdul-Mageed and Ungar (2017) collected about 250 million tweets pertaining to 24 emotions and 665 hashtags and developed a Gated Recurrent Neural Network (GRNN) model for emotion classification.

- **Work on Emotion-Labeled Datasets 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

---

18. http://emotion-research.net/toolbox/
    toolboxdatabase.2006-10-13.2581092615

19. www.experienceproject.com

with seven emotions: anger, disgust, fear, happiness, like, sadness and surprise.[20] 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). Recent works on Chinese emotion classification using transfer learning and neural networks include Yu, Marujo, Jiang, Karuturi, and Brendel (2018), Matsumoto, Sasayama, Yoshida, Kita, and Ren (2018).

Notable Arabic emotion datasets include the tweet datasets from the shared task Semeval-2018 task 1: Affect in Tweets Mohammad et al. (2018) and other works such as Badarneh, Al-Ayyoub, Alhindawi, Jararweh, et al. (2018), Al-Khatib and El-Beltagy (2017). Hussien, Tashtoush, Al-Ayyoub, and Al-Kabi (2016) analyzed the effectiveness of using emoticons for distant supervision in Arabic tweets.

- **Multimodal Datasets:** 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). Mohammad and Kiritchenko (2018) created a dataset of more than 4,000 pieces of art (mostly paintings) that has annotations for emotions evoked in the observer by the painting (image and title), the title alone, and the image alone. The pieces of art are selected from WikiArt.org's collection for four western styles (Renaissance Art, Post-Renaissance Art, Modern Art, and Contemporary Art). In addition, there is a large body of work on analyzing emotions through videos that analyzes facial expressions, gestures, text transcripts, and vocal features. (See survey by Marechal, Mikołajewski, Tyburek, Prokopowicz, Bougueroua, Ancourt, and Wegrzyn-Wolska (2019).) Such multi-modal representations of emotions are useful not only to determine the emotional states of entities more accurately but also in tasks such as captioning images or audio for emotions and even generating text that is affectually suitable for a given image or audio sequence.

As seen above, there are a number of datasets where sentences are manually labeled for emotions. They have helped improve our understanding of how people use language to convey emotions and they have helped in developing supervised machine learning emotion classification systems. Yet, a number of questions remain unexplored. For example, to what extent do people convey emotions without using explicit emotion words? are some emotions more basic than others? is there 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? and so on.

4.4 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 exam-

---

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

ple, 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, 2003b; Wiebe et al., 2004). Opinion Finder is a popular freely available subjectivity systems (Wilson, Hoffmann, Somasundaran, Kessler, Wiebe, Choi, Cardie, Riloff, & Patwardhan, 2005).[21] However, there has been very little work on subjectivity detection in the later half of 2010s; likely because of the greater interest in modeling sentiment and emotion directly as opposed to through the framework of subjectivity.

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

Even though some small but influential lexicons such as the General Inquirer (Stone et al., 1966) and ANEW (Bradley & Lang, 1999) were created decades earlier, the 2010s saw

---

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

notable progress in two salient areas: crowdsourced emotion lexicons and reliable real-valued annotations. We discuss the two in the sub-sections below, followed by a list of prominent emotion lexicons and their details.

5.1.1 Crowdsourced Emotion Lexicons

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.[22] 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 (EmoLex) was the first emotion lexicon created by crowdsourcing. It includes entries for ~14k English terms. Each entry includes ten binary scores (0 or 1) indicating no association or association with eight basic emotions as well as positive and negative sentiment (Mohammad & Turney, 2012, 2010).[23] For each word, annotators were first asked a simple word choice question where one of the options (the correct answer) is taken from a thesaurus, and the remaining options are made up of randomly selected words. This question acts as a mechanism to check whether the annotator knew the meaning of the word and was answering questions diligently. The question also serves a mechanism to bias the annotator response to a particular sense of the word. (Different senses of a word can convey differing emotions.) The resulting lexicon has entries for about 25K word senses. A word-level lexicon (for ~14K words) was created by taking the union of the emotions associated with each of the senses of a word.

Other examples of crowdsourced emotion lexicons include the NRC Valence, Arousal, and Dominance Lexicon (Mohammad, 2018a) and the NRC Affect Intensity Lexicon (Mohammad, 2018b). However, these required real-valued scores (and not just binary scores), which entail additional challenges. We discuss them next.

5.1.2 Real-valued Sentiment Scores from manual Annotations

Words have varying degrees 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

---

22. https://www.mturk.com/mturk/welcome
    http://www.crowdflower.com
23. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

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).

Best–worst scaling (BWS) is an annotation scheme that retains the comparative aspect of annotation while still requiring only a small number of annotations (Louviere, 1991; Louviere, Flynn, & Marley, 2015). It has its basis in the mathematical choice modeling and psycho-physics. Essentially, the annotator is presented with four (or five) items and asked which is the most (say, most positive) and which is the least (say, 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 we know:

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

Each of these BWS questions can be presented to multiple annotators.

The responses to the BWS 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). Kiritchenko and Mohammad (Kiritchenko & Mohammad, 2017) showed through empirical experiments that BWS produces more reliable and more discriminating scores than those obtained using rating scales.[24]

BWS 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, Zhu, and Mohammad (2014b) used BWS to create a dataset of 1500 Twitter terms with real-valued sentiment association scores.

Real-valued valence scores obtained through BWS 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 were used in a 2016 Task *Determining sentiment intensity of English and Arabic phrases* to evaluate both English and Arabic automatically generated sentiment lexicons.

BWS was also used to create the NRC Valence, Arousal, and Dominance Lexicon (NRC-VAD) (Mohammad, 2018a) and the NRC Emotion Intensity Lexicon (NRC-EIL) (Mohammad, 2018b). With over 20K entries, NRC-VAD is the largest manually created emotion lexicon. Mohammad (2018a) also analyzed the annotations to show how different demographic groups, such as men and women, people with different personality traits, etc., perceive valence, arousal, and dominance differently. NRC-EIL provides intensity scores for the words and emotions in the NRC Emotion Lexicon. Mohammad and Kiritchenko (2018) analyzed the entries across NRC-VAD and NRC-EIL to show that while the positive and negative emotions are clearly separated across the valence dimension, emotions such as anger, fear, and disgust do not occur in clearly separate regions of the VAD space.

---

24. See Kiritchenko and Mohammad (2016a, 2017) for further details on BWS and its use in NLP applications.

5.1.3 Large, Influential, Manually Created Lexicons

The earliest emotion lexicons such as the General Inquirer and ANEW capture words that denote (express) an emotion. These lexicons tend to be small. Starting with the NRC Emotion lexicon, we see work on capturing not just words that denote an emotion, but also those that connotate an emotion. For example, the word *party* does not denote joy, but is associated with, or connotes, joy. Below are some of the most widely used sentiment and emotion lexicons. They are ordered by size.

- **NRC Valence, Arousal, and Dominance Lexicon (NRC-VAD): ∼20K English terms** annotated for real-valued scores of **valence, arousal and dominance** (Mohammad, 2018a).[25] Automatic translations of the words in over 100 languages included.

- **NRC Emotion Lexicon (EmoLex): ∼14k English terms** with binary annotations (0 or 1) indicating no association or association with **eight basic emotions** (those proposed by Plutchik (1980)) as well as for positive and negative sentiment.[26] Automatic translations of the words in overl 100 languages included (Mohammad & Turney, 2010, 2012).

- **Warriner et al. Lexicon: ∼14K English terms** annotated for **valence, arousal, and dominance** (real-valued scores) (Warriner, Kuperman, & Brysbaert, 2013).[27]

- **NRC Emotion Intensity Lexicon: ∼10K English terms** annotated for intensity (real-valued) scores corresponding to **eight basic emotions** (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) (Mohammad, 2018b).[28]

- **MPQA Subjectivity Lexicon: ∼8,000 English terms** annotated for **valence** (strongly positive, weakly positive, strongly negative and weakly negative) (Wilson, Wiebe, & Hoffmann, 2005).[29] Includes the words and annotations from the General Inquirer and other sources.

- **Hu and Liu Lexicon: ∼6,800 English terms** from customer reviews that are annotated for **valence** (positive, negative) (Hu & Liu, 2004).[30]

- **General Inquirer (GI): ∼3,600 English terms** annotated for associations with various semantic categories including **valence** (Stone et al., 1966).[31] These include about 1500 words from the Osgood study.

- **AFINN: ∼2,500 English terms** annotated for **valence** (Nielsen, 2011). The ratings range from -5 (most negative) to +5 (most positive) in steps of 1.[32]

---

25. http://saifmohammad.com/WebPages/nrc-vad.html
26. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
27. http://crr.ugent.be/archives/1003
28. http://saifmohammad.com/WebPages/AffectIntensity.htm
29. https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
30. http://www.cs.uic.edu/∼liub/FBS/opinion-lexicon-English.rar
31. http://www.wjh.harvard.edu/ inquirer/
    Note: Stone et al. (1966) use the term evaluativeness instead of valence.
32. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
    Note: The publication describing AFINN does not state whether the name is an abbreviation.

- **Linguistic Inquiry and Word Count (LIWC) Dictionary: ∼1,400 English terms** manually identified to denote the **affect categories**: positive emotion, negative emotion, anxiety, anger, and sadness (Pennebaker, Boyd, Jordan, & Blackburn, 2015).[33]

- **The Affective Norms for English Words (ANEW): ∼1,000 English terms** annotated for **valence, arousal, and dominance** (Bradley & Lang, 1999).[34]

Emotion lexicons such as those listed above provide simple and effective means to analyze and draw inferences from large amounts of text. They are also used by machine learning systems to improve prediction accuracy, especially when the amount of training data is limited. Large emotion lexicons are also widely used in digital humanities, literary analyses, psychology, and art for a number purposes.

5.2 Automatically inducing term–sentiment association lexicons

Automatic methods for capturing word–sentiment associations can quickly learn associations for hundreds of thousands of words, and even for sequences of words. They can learn associations that are relevant to a particular domain (Chetviorkin, Moscow, & Loukachevitch, 2014; Hamilton, Clark, Leskovec, & Jurafsky, 2016). 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*). They also tend to have higher coverage (include more terms) than manually created lexicons. Notably, the coverage of domain specific terms is often much better in automatically generated lexicons than general purpose manually created lexicons. For example, a sentiment lexicon that is generated from a large number of freely available tweets, such as the Sentiment140 lexicon Mohammad et al. (2013a), has a number of entries for Twitter-specific languages such as emoticons (*:-)*), hashtags (*#love*), conjoined words (*loveumom*), and creatively spelled words (*yummeee*). Similarly, sentiment lexicons, such as SenTrop (Hamilton et al., 2016), can be generated that are appropriate for different historical time periods.[35] Automatic methods can also be used to create separate lexicons for words found in negated context and those found in affirmative context (Kiritchenko et al., 2014b); the idea being that the same word contributes to sentiment differently depending on whether it is negated or not. These lexicons can contain sentiment associations for hundreds of thousands of unigrams, and even larger textual units such as bigrams and trigrams.

Methods for automatically inducing sentiment lexicons rely on three basic components: a large lexical resource, a small set of of seed positive and negative terms (sometimes referred to as paradigm words), and a method to propagate or induce sentiment scores for the terms in the lexical resource. Commonly used lexical resources are: semantic networks such as WordNet (Hatzivassiloglou & McKeown, 1997; Baccianella, Esuli, & Sebastiani, 2010),

---

33. LIWC also includes other words associated with categories such as personal pronouns and biological concern. The LIWC dictionary also comes with software to analyze text and predict the psychological state of the writer by identifying repeated use of words from various categories.

34. http://csea.phhp.ufl.edu/media/anewmessage.html

35. Over time, words can change meaning and also sentiment. For example, words such as terrific and lean originally had negative connotations, but now they are largely considered to have positive connotations.

thesauri (Mohammad, Dunne, & Dorr, 2009), and Wikipedia (Chen & Skiena, 2014); text corpora such as the world wide web data (Turney & Littman, 2003) and collections of tweets Mohammad and Kiritchenko (2013), Abdul-Mageed and Ungar (2017). A popular variant in recent years is to make use of word embeddings (which themselves are generated from extremely large text corpora) as the primary lexical resource (Faruqui, Dodge, Jauhar, Dyer, Hovy, & Smith, 2015; Yu, Wang, Lai, & Zhang, 2017).

Seed words are compiled manually or through existing manually created sentiment lexicons. Turney and Littman (2003) and Hamilton et al. (2016) identified small lists of positive and negative words that are largely monosemous and have stable sentiments in different contexts and time periods. Go, Bhayani, and Huang (2009) used emoticons as seed terms to induce valence of words in tweets. Mohammad et al. (2013a) used hashtags (#angry, #sad, #good, #terrible, etc.) and emoticons (:), :() to induce valence and emotion association scores for eight basic emotions and positive negative sentiment. Mohammad and Kiritchenko (2013), Abdul-Mageed and Ungar (2017) used hashtags to induce emotion associations for dozens of different emotions.

The central idea in methods to induce sentiment lexicons is to propagate the sentiment information from the seed terms to other terms in the lexical resource. This method of propagation can make use of simple word–seed co-occurrence in text (Turney & Littman, 2003; Mohammad et al., 2013a) or any of the graph-propagation algorithms in a semantic network (Baccianella et al., 2010).

Word embeddings, which are induced from text corpora, are designed to produce word representations such that if two words are close in meaning, then their representations are close to each other (in vector space). They rely on the idea that words that are close in meaning also occur in similar contexts—distributional hypothesis (Firth, 1957; Harris, 1968). Methods to induce sentiment lexicons from word embeddings propagate sentiment labels or scores from the seed words to other words in the vector space in such a way that the sentiment score of a word is influenced strongly by the sentiment scores of the words closest to it. However, antonymous words are known to co-occur with each other more often than random chance (Cruse, 1986; Fellbaum, 1995; Mohammad, Dorr, & Dunn, 2008). Thus, to some extent, antonymous words tend to occur close to each other in word embeddings, making them somewhat problematic in the propagation of sentiment. Thus, several approaches have been proposed that make use of word embeddings (as the primary lexical resource) and some sentiment signal (star rating of reviews or seed set of sentiment words). These approaches are of two kinds: (1) those that learn word embeddings with an additional objective function to keep antonymous and differing sentiment word pairs farther away from each other (Maas, Daly, Pham, Huang, Ng, & Potts, 2011; Tang, Wei, Qin, Zhou, & Liu, 2014b; Ren, Zhang, Zhang, & Ji, 2016); and (2) those that modify (refine/retrofit) word embeddings post-hoc (Labutov & Lipson, 2013; Faruqui et al., 2015; Hamilton et al., 2016; Yu et al., 2017).

6. Modeling the impact of sentiment modifiers

Negation words (e.g., *no, not, never, hardly, can't, don't*), modal verbs (e.g., *can, may, should, must*), degree adverbs (e.g., *almost, nearly, seldom, very*) and other modifiers impact the sentiment of the term or phrase they modify. Kiritchenko and Mohammad (2016b)

manually annotated a set of phrases that include negators, modals, degree adverbs, and their combinations. Both the phrases and their constituent content words are annotated with real-valued scores of sentiment intensity using the technique BestWorst Scaling. They measured the effect of individual modifiers as well as the average effect of the groups of modifiers on overall sentiment. They found that the effect of modifiers varies substantially among the members of the same group. Furthermore, each individual modifier can affect the modified sentiment words in different ways. Their lexicon, *Sentiment Composition Lexicon of Negators, Modals, and Adverbs (SCL-NMA)*, was used as an official test set in the SemEval-2016 shared Task #7: Detecting Sentiment Intensity of English and Arabic Phrases. The objective of that task was to automatically predict sentiment intensity scores for multi-word phrases. Below we outline some focused work on individual groups of sentiment modifiers.

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 employed simple heuristics such as flipping the polarity of the words in a negator's scope (Kennedy & Inkpen, 2005) 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 Sarcasm, Metaphor, and other Figurative 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-Ibáñez, Muresan, & Wacholder, 2011). In 2015, a SemEval shared task was organized on detecting sentiment in tweets rich in metaphor and irony (Task 11).[36] 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 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.

It is generally believed that a metaphor tends to have an emotional impact, and thus it is not surprising that they are used widely in language. However, their inherent non-literality can pose a challenge to sentiment analysis systems. Further, the mechanisms

---

36. The proceedings will be released later in 2015.

through which metaphors convey emotions are not well understood. Mohammad, Shutova, and Turney (2016) presented a study comparing the emotionality of metaphorical English expressions with that of their literal counterparts. They found that metaphorical usages are, on average, significantly more emotional than literal usages. They also showed that this emotional content is not simply transferred from the source domain into the target, but rather is a result of meaning composition and interaction of the two domains in the metaphor. Ho and Cheng (2016) analyses a text of English financial analysis reports for the use of emotional metaphors. Rai, Chakraverty, Tayal, Sharma, and Garg (2019) show how models for understanding metaphors benefit from sensing the associated emotions. See work by (Dankers, Rei, Lewis, & Shutova, 2019) on joint models for detecting metaphor and sentiment. (Su, Li, Peng, & Chen, 2019) developed a method to detect sentiment of Chinese metaphors.

Balahur, Steinberger, Kabadjov, Zavarella, der Goot, Halkia, Pouliquen, and Belyaeva (2010) and Liu, Qian, Qiu, and Huang (2017) showed that automatic sentiment analysis systems often have difficulties in dealing with idioms. Williams, Bannister, Arribas-Ayllon, Preece, and Spasić (2015) and Spasic, Williams, and Buerki (2017) show that sentiment analysis in English texts can be improved using idiom-related features. Williams et al. (2015) and Jochim, Bonin, Bar-Haim, and Slonim (2018) annotated lists of 580 and 5000 English idioms, respectively, that are manually annotated for sentiment. Passaro, Senaldi, and Lenci (2019) collected valence and arousal ratings for 45 Italian verb-noun idioms and 45 Italian non-idiomatic verb-noun pairs.

We found little to no work exploring automatic sentiment detection in hyperbole, understatement, rhetorical questions, and other creative uses of language.

8. Multilingual and Crosslingual Sentiment Analysis

A large proportion of research in sentiment analysis, and natural language processing in general, has focused on English. Thus for languages other than English there are fewer and smaller resources (sentiment lexicons, emotion-annotated corpora, etc.). This means that automatic sentiment analysis systems in other languages tend to be less accurate than their English counterpart. Work in *Multilingual Sentiment Analysis* aims at building multilingual affect-related resources, as well as developing (somewhat) language independent approaches for sentiment analysis—approaches that can be applied to wide variety of resource-poor languages. A common approach to sentiment analysis in a resource-poor *target language* is to leverage powerful resources from another *source language* (usually English) via translation. This type of work is usually referred to as *Crosslingual Sentiment Analysis*.

Work in multilingual and crosslingual sentiment analysis includes that on leveraging source language sentiment annotated corpora (Pan, Xue, Yu, & Wang, 2011; Salameh, Mohammad, & Kiritchenko, 2015; Chen, Sun, Athiwaratkun, Cardie, & Weinberger, 2018); unlabeled bilingual parallel data (Meng, Wei, Liu, Zhou, Xu, & Wang, 2012), source language sentiment lexicons (Mihalcea, Banea, & Wiebe, 2007; Mohammad, Salameh, & Kiritchenko, 2015), multilingual WordNet (Bobicev, Maxim, Prodan, Burciu, & Anghelus, 2010; Cruz, Troyano, Pontes, & Ortega, 2014), multilingual word and sentiment embeddings (learned from sentiment annotated data in source and target language) (Zhou, Wan, & Xiao, 2016; Feng & Wan, 2019), etc.

Duh, Fujino, and Nagata (2011) presented an opinion piece on the challenges and opportunities of using automatic machine translation for crosslingual sentiment analysis, and notably suggest the need for new domain adaptation techniques. Balahur and Turchi (2014) conducted a study to assess the performance of sentiment analysis techniques on machine-translated texts. Salameh et al. (2015) and Mohammad et al. (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. They also showed that automatic Arabic translations of English valence lexicons improve accuracies of an Arabic sentiment analysis system.[37]

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. Applications

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

- *Commerce, Brand Management, Customer Relationship Management, and FinTech (Finance and Technology)*: 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).

- *Art*: Art and emotions are known to have a strong connection—art often tends to be evocative. Thus it is not surprising to see the growing use of emotion resources in art. (Davis & Mohammad, 2014b) developed a syste, *TransProse*, that generates music that captures the emotions in a piece of literature. The system uses the NRC Emotion Lexicon. A symphony orchestra performed music composed using TransProse and a human composer at the Louvre museum in Paris, September 20, 2016. Wishing Wall, an interactive art installation that visualizes wishes using the NRC Emotion lexicon was displayed in Tekniska Museet, Stockholm, Sweden (Oct 2014–Aug 2015), Onassis Cultural Centre, Athens (Oct 2015–Jan 2016), and Zorlu Centre in Istanbul (Feb 12th–June16). WikiArt Emotions that annotated emotions evoked by art (Mohammad & Kiritchenko, 2018).

- *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).

---

37. The translated lexica and corpora for Arabic sentiment analysis: www.purl.org/net/ArabicSA

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).

- *Literary Analysis and Digital Humanities*: 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 (Mohammad, 2011, 2012b; Hartner, 2013). There is also work in generating music that captures the emotions in text (Davis & Mohammad, 2014a).

- *Personality Traits*: Systematic patterns in how people express emotions are key indicators 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; Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, Agrawal, Park, et al., 2013; Mohammad & Kiritchenko, 2013).

- *Understanding Gender Differences*: Men and women 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 (Mohammad & Yang, 2011b; Grijalva et al., 2014; Montero, Munezero, & Kakkonen, 2014).

- *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, 2011). Tweet streams have been shown to help identify current public opinion towards the candidates in an election (nowcasting) (Golbeck & Hansen, 2011; Mohammad, Zhu, Kiritchenko, & Martin, 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, 2010; Lampos, Preotiuc-Pietro, & Cohn, 2013). However, other research expresses skepticism at the extent to which forecasting is possible (Avello, 2012).

- *Public Health and Psychology*: Automatic methods for detecting emotions are useful in detecting depression (Pennebaker, Mehl, & Niederhoffer, 2003; Cherry, Mohammad, & De Bruijn, 2012), identifying cases of cyber-bullying (Chen, Zhou, Zhu, & Xu, 2012), predicting health attributes at community level (Johnsen, Vambheim,

Wynn, & Wangberg, 2014; Eichstaedt, Schwartz, Kern, Park, Labarthe, Merchant, Jha, Agrawal, Dziurzynski, Sap, et al., 2015), gender differences in how we perceive connotative word meaning (Mohammad, 2018a), and tracking well-being (Schwartz 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.

- *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) 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.

10. Ethics and Fairness

Recent advances in machine learning have meant that computer-aided systems are becoming more human-like in their predictions. This also means that they perpetuate human biases. Some learned biases may be beneficial for the downstream application. Other biases can be inappropriate and result in negative experiences for some users. Examples include, loan eligibility and crime recidivism systems that negatively assess people belonging to a certain area code (which may disproportionately impact people of a certain race) (Chouldechova, 2017) and resumé sorting systems that believe that men are more qualified to be programmers than women (Bolukbasi et al., 2016). Similarly, sentiment and emotion analysis systems can also perpetuate and accentuate inappropriate human biases. In fact, recent work has shown that a large majority of sentiment and emotion analysis machine learning systems consistently give different emotionality scores to sentences mentioning different races and genders (Kiritchenko and Mohammad, 2018). Often these biases are across stereotypical lines, for example, marking near-identical sentences that mention African Americans names as being more angry than sentences that mention European American names, and sentences that mention women as being more emotional (more happy / more sad) than sentences that mention men. Thelwall (2018) found that automatic sentiment analysis systems are somewhat less accurate on customer reviews written by men compared to those written by women. Díaz, Johnson, Lazar, Piper, and Gergle (2018) examined several sentiment resources, including word embeddings, to find that several old-age-related terms (and their mentions in text) are marked as having negative sentiment whereas terms associated with the young (and their mentions in text) are often evaluated as being positive. Thus sentiment analysis in the real-world has potential negative implications for the elderly.

Sentiment analysis, like other natural language technologies, can be a *great enabler*—allowing one to capitalize on substantial, previously unattainable, opportunities. However,

the opportunities can also be used by bad actors for committing maleficence. For example, imagine a world where sentiment analysis of social media posts can easily lead to determining one's position on a wide variety of issues, likes and dislikes, personality traits, as well as the emotional state when most pliable for persuasion (say, to purchase items online or to support a political agenda). Suddenly, we (both as individuals and whole populations) are open for manipulation, deception, and indoctrination. Unfortunately, that world is already upon us. To name just two such instances from the recent past: Facebook was reported to have told advertisers that it can identify teenagers who "feel stressed, "defeated, "overwhelmed, "anxious, "nervous, "stupid, "silly, "useless and a "failure.[38] Cambridge Analytica was accused to illegally extracting profiles of millions of Facebook users and leverage it to sway public opinion in the 2016 US elections (Cadwalladr & Graham-Harrison, 2018). Thus, despite the many benefits of sentiment analysis (such as those stated in the previous section), both the researcher and the lay person have to be on guard for the perils it will inevitably germinate.

11. 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 work preliminary work on sentiment modifiers such as negators and modals, work on 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.

Natural Language Processing (NLP), the broader field that encompasses sentiment analysis, has always been an interdisciplinary field with strong influences from Computer Science, Linguistics, and Information Sciences. However, over the last decade, NLP and Sentiment Analysis have made marked inroads into other fields of study such as psychology, digital humanities, history, art, and social sciences—sometimes attracting controversy. (For example, see the articles on the so called Digital Humanities Wars.[39]) A common principle in these advances is that NLP allows for asking old and new questions in these fields by examining massive amounts of text. Over the coming decade, we expect more of this, with the traditional boundaries between NLP and other fields becoming even more blurry, along with a healthy cross-pollination of ideas. Finally, we also expect to see considerable interest in the ethics, fairness, and bias of sentiment analysis systems, as well as the use of sentiment analysis to help identify explicit and implicit biases of people.

---

38. https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens
39. https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986

References

Abdou, M., Kulmizev, A., & Ginés i Ametllé, J. (2018). AffecThor at SemEval-2018 task 1: A cross-linguistic approach to sentiment intensity quantification in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 210–217, New Orleans, Louisiana. Association for Computational Linguistics.

Abdul-Mageed, M., & Ungar, L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 718–728, Vancouver, Canada. Association for Computational Linguistics.

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.

Agrawal, P., & Suri, A. (2019). NELEC at SemEval-2019 task 3: Think twice before going deep. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 266–271, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information processing & management*, *56*(2), 320–342.

Al-Khatib, A., & El-Beltagy, S. R. (2017). Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 105–114. Springer.

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.

Alowisheq, A., Alhumoud, S., Altwairesh, N., & Albuhairi, T. (2016). Arabic sentiment analysis resources: a survey. In *Social Computing and Social Media: 8th International Conference, SCSM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17–22, 2016. Proceedings 8*, pp. 267–278. Springer.

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.

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*.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceeding of the 7th International Conference on Language Resources and Evaluation*, Vol. 10 of *LREC '10*, pp. 2200–2204.

Badarneh, O., Al-Ayyoub, M., Alhindawi, N., Jararweh, Y., et al. (2018). Fine-grained emotion analysis of arabic tweets: A multi-target multi-label approach. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 340–345. IEEE.

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

Balahur, A., Steinberger, R., Kabadjov, M. A., Zavarella, V., der Goot, E. V., Halkia, M., Pouliquen, B., & Belyaeva, J. (2010). Sentiment analysis in the news. *ArXiv*, *abs/1309.6202*.

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.

Berger, J. (2013). *Contagious: Why things catch on.* Simon and Schuster.

Bifet, A., Holmes, G., Pfahringer, B., & Gavalda, R. (2011). Detecting sentiment change in twitter streaming data..

Bobicev, V., Maxim, V., Prodan, T., Burciu, N., & Anghelus, V. (2010). Emotions in words: Developing a multilingual wordnet-affect. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 375–384. Springer.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

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

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 4349–4357.

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.

Boumaiza, A. (2015). A survey on sentiment analysis and visualization. *Journal of Emerging Technologies in Web Intelligence Vol, 7*(1).

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.

Cabrio, E., & Villata, S. (2018). Five years of argument mining: a data-driven analysis.. In *IJCAI*, pp. 5427–5433.

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian, 17*, 22.

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.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chen, P., Sun, Z., Bing, L., & Yang, W. (2017). Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 452–461, Copenhagen, Denmark. Association for Computational Linguistics.

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., & Weinberger, K. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics, 6*, 557–570.

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

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.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153–163.

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. (2011). Political polarization on Twitter. *Networks*, *133*(26), 89–96.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge University Press.

Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, *41*(13), 5984–5994.

Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., & Duan, P. (2016). Word embeddings and convolutional neural network for Arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2418–2427, Osaka, Japan. The COLING 2016 Organizing Committee.

Dankers, V., Rei, M., Lewis, M., & Shutova, E. (2019). Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2218–2229.

Darwish, K., Stefanov, P., Aupetit, M. J., & Nakov, P. (2019). Unsupervised user stance detection on twitter. *arXiv preprint arXiv:1904.02000.*

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

Davis, H., & Mohammad, S. (2014b). 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.*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota.

Díaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Do, H. H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, *118*, 272–299.

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

Duh, K., Fujino, A., & Nagata, M. (2011). Is machine translation ripe for cross-lingual sentiment classification?. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 429–433, Portland, Oregon, USA. Association for Computational Linguistics.

Duppada, V., Jain, R., & Hiray, S. (2018). SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 18–23, New Orleans, Louisiana. Association for Computational Linguistics.

Ebrahimi, J., Dou, D., & Lowd, D. (2016). A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2656–2665.

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.*

Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pp. 48–54, Austin, TX, USA. Association for Computational Linguistics.

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.

El-Naggar, N., El-Sonbaty, Y., & El-Nasr, M. A. (2017). Sentiment analysis of modern standard arabic and egyptian dialectal arabic tweets. In *2017 Computing Conference*, pp. 880–887. IEEE.

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.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Fellbaum, C. (1995). Co-occurrence and antonymy. *International Journal of Lexicography*, *8*, 281–303.

Feng, Y., & Wan, X. (2019). Learning bilingual sentiment-specific word embeddings without cross-lingual supervision. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 420–429.

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. *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, *280*(1), 20–32.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55.. In *Studies in Linguistic Analysis*, pp. 1–32, Oxford: The Philological Society. (Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, Longman).

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.

Fraser, K. C., Zeller, F., Smith, D. H., Mohammad, S. M., & Rudzicz, F. (2019). How do we feel when a robot dies? emotions expressed on twitter before and after hitchbots destruction. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 62–71.

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

Gamal, D., Alfonse, M., El-Horbaty, E.-S. M., & Salem, A.-B. M. (2019). Twitter benchmark dataset for arabic sentiment analysis. *Int J Mod Educ Comput Sci*, *11*(1), 33.

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*.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. In *Technical report, Stanford University*.

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.

Gregory, M. L., Chinchor, N., Whitney, P., Carter, R., Hetzler, E., & Turner, A. (2006). User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pp. 23–30. 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...

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2016, p. 595.

Harris, Z. (1968). *Mathematical Structures of Language*. Interscience Publishers.

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.

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*.

Ho, J., & Cheng, W. (2016). Metaphors in financial analysis reports: How are emotions expressed?. *English for Specific Purposes*, *43*, 37–48.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

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.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the j-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144–147.

Huang, C., Trabelsi, A., & Zaïane, O. (2019). ANA at SemEval-2019 task 3: Contextual emotion detection in conversations through hierarchical LSTMs and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 49–53, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hussien, W. A., Tashtoush, Y. M., Al-Ayyoub, M., & Al-Kabi, M. N. (2016). Are emoticons good enough to train emotion classifiers of arabic tweets?. In *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, pp. 1–6. IEEE.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media.*

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.

Jochim, C., Bonin, F., Bar-Haim, R., & Slonim, N. (2018). SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

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.

Jordan, A. H., Monin, B., Dweck, C. S., Lovett, B. J., John, O. P., & Gross, J. J. (2011). Misery has more company than people think: Underestimating the prevalence of others negative emotions. *Personality and Social Psychology Bulletin*, *37*(1), 120–135.

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. H.-J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of ebola virus on twitter and in the news. *Journal of Information Science*, *42*(6), 763–781.

Kiritchenko, S., & Mohammad, S. M. (2016a). Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Kiritchenko, S., & Mohammad, S. M. (2016b). The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*.

Kiritchenko, S., & Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of The Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

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*.

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.

Kucher, K., Paradis, C., & Kerren, A. (2018). The state of the art in sentiment visualization. In *Computer Graphics Forum*, Vol. 37, pp. 71–96. Wiley Online Library.

Kücük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, *53*(1), 1–37.

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.

Labutov, I., & Lipson, H. (2013). Re-embedding words. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 489–493, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Laskari, N. K., & Sanampudi, S. K. (2016). Aspect based sentiment analysis survey. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *18*(2), 24–28.

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

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, *45*(4), 765–818.

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., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.

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.

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, P., Gulla, J. A., & Zhang, L. (2016). Dynamic topic-based sentiment analysis of large-scale online news. In *International Conference on Web Information Systems Engineering*, pp. 3–18. Springer.

Liu, P., Qian, K., Qiu, X., & Huang, X. (2017). Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1204–1213, Copenhagen, Denmark.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

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.

Lu, Y., Hu, X., Wang, F., Kumar, S., Liu, H., & Maciejewski, R. (2015). Visualizing social media sentiment in disaster scenarios. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1211–1215.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Marechal, C., Mikołajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., & Wegrzyn-Wolska, K. (2019). Survey on ai-based multimodal methods for emotion detection. In *High-Performance Modelling and Simulation for Big Data Applications*, pp. 307–324. Springer.

Matsumoto, K., Sasayama, M., Yoshida, M., Kita, K., & Ren, F. (2018). Transfer learning based on utterance emotion corpus for lyric emotion estimation. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 699–703. IEEE.

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)..

Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., & Wang, H. (2012). Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 572–581, Jeju Island, Korea.

Meshi, D., Morawetz, C., & Heekeren, H. R. (2013). Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use. *Frontiers in human neuroscience*, *7*, 439.

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., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

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. (2018). WikiArt emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41.

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.

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. Association for Computational Linguistics.

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. (2018a). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

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

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

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

Mohammad, S. M., Dorr, B., & Dunn, C. (2008). Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pp. 982–991, Waikiki, Hawaii.

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., & Kiritchenko, S. (2018). Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). 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. M., Salameh, M., & Kiritchenko, S. (2015). How translation alters sentiment. In *Journal of Artificial Intelligence Research*.

Mohammad, S. M., Shutova, E., & Turney, P. D. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*, Berlin, Germany.

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, *17*(3), 1–23.

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.

Morante, R., & Sporleder, C. (2012). Modality and negation: An introduction to the special issue. *Computational linguistics*, *38*(2), 223–260.

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. (2013). Character-to-character sentiment analysis in shakespeares plays..

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..

Oatley, K., & Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and emotion*, *1*(1), 29–50.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. F. (2009). Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pp. 9–16.

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.

Pan, J., Xue, G.-R., Yu, Y., & Wang, Y. (2011). Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In *PAKDD*.

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

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86. Association for Computational Linguistics.

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

Passaro, L. C., Senaldi, M. S., & Lenci, A. (2019). Do idioms have a heart? the side (sentiment of idiomatic expressions) project. *Education and Research Archive (ERA)*.

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

Peng, H., Cambria, E., & Hussain, A. (2017). A review of sentiment analysis research in chinese language. *Cognitive Computation, 9*(4), 423–435.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of liwc2015. Tech. rep..

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.

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

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana.

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

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.

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., & Manand-har, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Pool, C., & Nissim, M. (2016). Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pp. 30–39, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

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

Rai, S., Chakraverty, S., Tayal, D. K., Sharma, D., & Garg, A. (2019). Understanding metaphors using emotions. *New Generation Computing*, *37*(1), 5–27.

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.

Ren, Y., Zhang, Y., Zhang, M., & Ji, D.-H. (2016). Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings. In *AAAI*.

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. (2003a). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 105–112.

Riloff, E., & Wiebe, J. (2003b). 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.

Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 999–1005, Austin, Texas. Association for Computational Linguistics.

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.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813–830.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464.

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*.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, *285*, 181–203.

Sobhani, P., Mohammad, S., & Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pp. 159–169.

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.

Spasic, I., Williams, L., & Buerki, A. (2017). Idiombased features in sentiment analysis: Cutting the gordian knot. *IEEE Transactions on Affective Computing*.

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.

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

Su, C., Li, J., Peng, Y., & Chen, Y. (2019). Chinese metaphor sentiment computing via considering culture. *Neurocomputing*, *352*, 33–41.

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.

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., Qin, B., Feng, X., & Liu, T. (2016). Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Taulé, M., Martí, M. A., Rangel, F. M., Rosso, P., Bosco, C., Patti, V., et al. (2017). Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, Vol. 1881, pp. 157–177. CEUR-WS.

Thelwall, M. (2018). Gender bias in sentiment analysis. *Online Information Review*.

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

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

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.

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

Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Vohra, M. S., & Teraiya, J. (2013). Applications and challenges for sentiment analysis: A survey. *International journal of engineering research and technology*, *2*(2), 1–6.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

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 2008 Conference on Empirical Methods in Natural Language Processing*, 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, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2016). Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 225–230, Berlin, Germany. Association for Computational Linguistics.

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, 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.

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.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 246–253, College Park, Maryland, USA. Association for Computational Linguistics.

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.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A. D., & Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Syst. Appl.*, *42*, 7375–7385.

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.

Xiao, Z., Li, X., Wang, L., Yang, Q., Du, J., & Sangaiah, A. K. (2018). Using convolution control block for chinese sentiment analysis. *Journal of Parallel and Distributed Computing*, *116*, 18–26.

Xu, B., Mohtarami, M., & Glass, J. (2019). Adversarial domain adaptation for stance detection. *arXiv preprint arXiv:1902.02401*.

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.

Yan, Y., Chen, J., & Shyu, M.-L. (2020). Efficient large-scale stance detection in tweets. In *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications*, pp. 667–683. IGI Global.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding.. pp. 5754–5764.

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

Yin, D., Meng, T., & Chang, K.-W. (2020). Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association of Computational Linguistics (ACL-2020)*, Seattle, USA.

Yu, J., Marujo, L., Jiang, J., Karuturi, P., & Brendel, W. (2018). Improving multi-label emotion classification via sentiment classification with dual attention transfer network.. ACL.

Yu, L.-C., Wang, J., Lai, K. R., & Zhang, X. (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 534–539, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Zhou, J., Lu, Y., Dai, H.-N., Wang, H., & Xiao, H. (2019). Sentiment analysis of chinese microblog based on stacked bidirectional lstm. *IEEE Access*, *7*, 38856–38866.

Zhou, X., Wan, X., & Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1403–1412, Berlin, Germany.

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*.