Challenges in Sentiment Analysis

Saif M. Mohammad Saif.Mohammad@nrc-cnrc.gc.ca
National Research Council Canada
1200 Montreal Rd., Ottawa, ON, Canada

1. Introduction

There has been a large volume of work in sentiment analysis over the past decade and it continues to rapidly develop in new directions. A vast majority of this work has been on developing more accurate sentiment classifiers, usually involving supervised machine learning algorithms and a battery of features. Surveys by Pang and Lee (2008), Liu and Zhang (2012), and Mohammad (2016b) give summaries of the many automatic classifiers, features, and datasets used to detect sentiment. In this chapter, we flesh out some of the challenges that still remain, questions that have not been explored sufficiently, and new issues emerging from taking on new sentiment analysis problems. We also discuss proposals to deal with these challenges. The goal of this chapter is to equip researchers and practitioners with pointers to the latest developments in sentiment analysis and encourage more work in the diverse landscape of problems, especially those areas that are relatively less explored.

We start in Section 2 by discussing different sentiment analysis problems and how one of the challenges is to explore new sentiment analysis problems that go beyond simply determining whether a piece of text is positive, negative, or neutral. Some of the more ambitious problems that need more work include detecting sentiment at various levels of text granularities (terms, sentences, paragraphs, etc); detecting sentiment of the reader or sentiment of entities mentioned in the text; detecting sentiment towards aspects of products; detecting stance towards pre-specified targets that may not be explicitly mentioned in the text and that may not be the targets of opinion in the text; and detecting semantic roles of sentiment. Since many sentiment analysis systems rely on sentiment lexicons, we discuss capabilities and limitations of existing manually and automatically created sentiment lexicons in Section 3. In Section 4, we discuss the difficult problem of sentiment composition—how to predict the sentiment of a combination of terms. More specifically, we discuss the determination of sentiment of phrases (that may include negators, degree adverbs, and intensifiers) and sentiment of sentences and tweets. In Section 5, we discuss challenges in annotation of data for sentiment. We provide categories of sentences that are particularly challenging for sentiment annotation. Section 6 presents challenges in multilingual sentiment analysis. This is followed by a discussion on the challenges of applying sentiment analysis to downstream applications, and finally, some concluding remarks (Section 7).

2. The Array of Sentiment Analysis Tasks

Sentiment analysis is a generic name for a large number of opinion and affect related tasks, each of which present their own unique challenges. The sub-sections below provide an overview.

2.1 Sentiment at Different Text Granularities

Sentiment can be determined at various levels: from sentiment associations of words and phrases; to sentiment of sentences, SMS messages, chat messages, and tweets; to sentiment in product reviews, blog posts, and whole documents. A word–sentiment (or valence) association lexicon may have entries such as:

delighted – positive
killed – negative
shout – negative
desk – neutral

These lexicons can be created either by manual annotation or through automatic means. Manually created lexicons tend to be in the order of a few thousand entries, but automatically generated lexicons can capture sentiment 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). Entries in an automatically generated lexicon often also include a real-valued score indicating the strength of association between the word and the valence category. These numbers are prior estimates of the sentiment of terms in an average usage of the term. While sentiment lexicons are often useful in sentence-level sentiment analysis[1], the same terms may convey different sentiments in different contexts. The SemEval 2013 and 2014 *Sentiment Analysis in Twitter* shared tasks had a separate sub-task aimed at identifying sentiment of terms in context. Automatic systems have largely performed well in this task, obtaining F-scores close to 0.9. We discuss manually and automatically created sentiment lexicons in more detail in Section 3.

Sentence-level valence classification systems assign labels such as positive, negative, or neutral to whole sentences. It should be noted that the valence of a sentence is not simply the sum of the polarities of its constituent words. Automatic systems learn a model from labeled training data (instances that are already marked as positive, negative, or neutral) using a large number of features such as word and character ngrams, valence association lexicons, negation lists, word clusters, and even embeddings-based features. In recent years, there have been a number of shared task competitions on valence classification such as the 2013, 2014, and 2015 SemEval shared tasks titled *Sentiment Analysis in Twitter*, the 2014 and 2015 SemEval shared tasks on *Aspect Based Sentiment Analysis*, the 2015 SemEval shared task *Sentiment Analysis of Figurative Language in Twitter*, and the 2015 Kaggle competition *Sentiment Analysis on Movie Reviews*.[2] The NRC-Canada system (Mohammad, Kiritchenko, & Zhu, 2013a; Kiritchenko, Zhu, & Mohammad, 2014b), a supervised

1. The top systems in the SemEval-2013 and 2014 *Sentiment Analysis in Twitter* tasks used large sentiment lexicons (Wilson, Kozareva, Nakov, Rosenthal, Stoyanov, & Ritter, 2013; Rosenthal, Nakov, Ritter, & Stoyanov, 2014a).
2. http://alt.qcri.org/semeval2015/task10/
http://alt.qcri.org/semeval2015/task12/

machine learning system, came first in the 2013 and 2014 competitions. Other sentiment analysis systems developed specifically for tweets include those by Pak and Paroubek (2010), Agarwal, Xie, Vovsha, Rambow, and Passonneau (2011), Thelwall, Buckley, and Paltoglou (2011), Brody and Diakopoulos (2011), Aisopos, Papadakis, Tserpes, and Varvarigou (2012), Bakliwal, Arora, Madhappan, Kapre, Singh, and Varma (2012). However, even the best systems currently obtain an F-score of only about 0.7.

Sentiment analysis involving many sentences is often broken down into the sentiment analysis of the component sentences. However, there is interesting work in sentiment analysis of documents to generate text summaries (Ku, Liang, & Chen, 2006; Liu, Cao, Lin, Huang, & Zhou, 2007; Somprasertsri & Lalitrojwong, 2010; Stoyanov & Cardie, 2006; Lloret, Balahur, Palomar, & Montoyo,), as well as detecting the patterns of sentiment and detecting sentiment networks in novels and fairy tales (Nalisnick & Baird, 2013b, 2013a; Mohammad & Yang, 2011).

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

On the surface, sentiment may seem unambiguous, but looking closer, it is easy to see how 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. A large majority of 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:

James: *The pop star suffered a fatal overdose of heroine.*

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

On the other hand, people can react differently to the same utterance, for example, people on opposite sides of a debate or rival sports fans. Thus modeling listener sentiment requires modeling listener profiles. This is an area of research not explored much by the community. Similarly, there is no work on modeling sentiment of entities mentioned in the text, for example, given:

Drew: *Jackson could not stop talking about the new Game of Thrones episode.*

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

http://alt.qcri.org/semeval2015/task11/
http://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

2.3 Sentiment Towards Aspects of an Entity

A review of a product or service can express sentiment towards various aspects. For example, a restaurant review can speak positively about the service, but express a negative attitude towards the food. There is now a growing amount of work in detecting aspects of products in text and also in determining sentiment towards these aspects. In 2014, a shared task was organized for detecting aspect sentiment in restaurant and laptop reviews (Pontiki, Galanis, Pavlopoulos, Papageorgiou, Androutsopoulos, & Manandhar, 2014a). The best performing systems had a strong sentence-level sentiment analysis system to which they added localization features so that more weight was given to sentiment features close to the mention of the aspect. This task was repeated in 2015. It will be useful to develop aspect-based sentiment systems for other domains such as blogs and news articles as well. (See proceeding of SemEval-2014 and 2015 for details about participating aspect sentiment systems.)

2.4 Stance Detection

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

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

Humans can deduce from the text that the speaker is against the proposition. However, this is a challenging task for computers. To successfully detect stance, automatic systems often have to identify relevant bits of information that may not be present in the focus text. The systems also have to first identify the target of opinion in the text and then determine its implication on the target of interest. Note that the target of opinion need not be the same as the target of interest. For example, that if one is actively supporting foetus rights (target of opinion), then he or she is likely against the right to abortion (target of interest). Automatic systems can obtain such information from large amounts of domain text.

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

Mohammad, Sobhani, and Kiritchenko (2016b) created the first dataset of tweets labeled for both stance and sentiment. More than 4000 tweets are annotated for whether one can deduce favorable or unfavorable stance towards one of five targets 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', and 'Legalization of Abortion'. Each of these tweets is also annotated for whether the target of opinion expressed in the tweet is the same as the given target of interest. Finally, each tweet is annotated for whether it conveys positive, negative, or neutral sentiment. Partitions of this stance-annotated data were used as training and test sets in the SemEval-2016 shared task competition, Task #6: Detecting Stance from Tweets (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016a). Participants were provided with 2,914 training instances labeled for stance for the five targets. The test data included 1,249 instances. All of the stance data

Role	Description
Core:	
Emotion	the feeling that the Experiencer experiences
State	the state the Experiencer is in
Evaluation	a negative or positive assessment of the Experiencer regarding his/her State
Experiencer	one who experiences the Emotion and is in the State.
Non-Core:	
Explanation	The thing that leads to the Experiencer feeling the Emotion or State.

Table 1: The FrameNet frame for feeling.

is made freely available through the shared task website. The task received submissions from 19 teams. The best performing system obtained an overall average F-score of 67.8 in a three-way classification task: favour, against, or neither. They employed two recurrent neural network (RNN) classifiers: the first was trained to predict task-relevant hashtags on a large unlabeled Twitter corpus. This network was used to initialize a second RNN classifier, which was trained with the provided training data (Zarrella & Marsh, 2016). Mohammad et al. (2016b) developed a SVM system that only uses features drawn from word and character ngrams and word embeddings to obtain an even better F-score of 70.3 on the shared task's test set. Yet, performance of systems is substantially lower on tweets where the target of opinion is an entity other than the target of interest.

Most of the earlier work focused on two-sided debates, for example on congressional debates (Thomas et al., 2006) or debates in online forums (Somasundaran and Wiebe, 2009; Murakami and Raymond, 2010; Anand et al., 2011; Walker et al., 2012; Hasan and Ng, 2013; Sridhar, Getoor, and Walker, 2014). New research in domains such as social media texts, and approaches that combine traditional sentiment analysis with relation extraction can make a significant impact in improving the state-of-the-art in automatic stance detection.

2.5 Detecting Semantic Roles of Feeling

Past work in sentiment analysis has focused extensively on detecting polarity, and to a smaller extent on detecting the target of the sentiment (the stimulus) (Popescu & Etzioni, 2005; Su, Xiang, Wang, Sun, & Yu, 2006; Xu, Huang, & Wang, 2013; Qadir, 2009; Zhang, Liu, Lim, & O'Brien-Strain, 2010; Zhang & Liu, 2011; Kessler & Nicolov, 2009). However, there exist other aspects relevant to sentiment. Tables 1 and 2 show FrameNet (Baker, Fillmore, & Lowe, 1998) frames for 'feelings' and 'emotions', respectively. Observe that in addition to Evaluation, State, and Stimulus, several other roles such as Reason, Degree, Topic, and Circumstance are also of significance and beneficial to down-stream applications such as information retrieval, summarization, and textual entailment. Detecting these various roles is essentially a semantic role-labeling problem (Gildea & Jurafsky, 2002; Màrquez, Carreras, Litkowski, & Stevenson, 2008; Palmer, Gildea, & Xue, 2010), and it is possible that they can be modeled jointly to improve detection accuracy. Li and Xu (2014) proposed

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 in the Experiencer.
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 a certain emotional response

Table 2: The FrameNet frame for emotions.

a rule-based system to extract the event that was the cause of an emotional Weibo (Chinese microblogging service) message. Mohammad, Zhu, Kiritchenko, and Martin (2015a) created a corpus of tweets from the run up to the 2012 US presidential elections, with annotations for sentiment, emotion, stimulus, and experiencer. The data also includes annotations for whether the tweet is sarcastic, ironic, or hyperbolic. Diman Ghazi and Szpakowicz (2015) compiled FrameNet sentences that were tagged with the stimulus of certain emotions.

2.6 Detecting Affect and Emotions

Sentiment analysis is most commonly used to refer to the goal of determining the valence or polarity of a piece of text. However, it can refer more generally to determining one's attitude towards a particular target or topic. Here, attitude can even mean emotional or affectual attitude such as frustration, joy, anger, sadness, excitement, and so on. Russell (1980) developed a circumplex model of affect and showed that it can be characterized by two primary dimensions: valence (positive and negative dimension) and arousal (degree of reactivity to stimulus). Thus, it is not surprising that large amounts of work in sentiment analysis is focused on determining valence. However, there is barely any work on automatically detecting arousal and a relatively small amount of work on detecting emotions such as anger, frustration, sadness, and optimism (Strapparava & Mihalcea, 2007; Aman & Szpakowicz, 2007; Tokuhisa, Inui, & Matsumoto, 2008; Neviarouskaya, Prendinger, & Ishizuka, 2009; Bellegarda, 2010; Mohammad, 2012; Boucouvalas, 2002; Zhe & Boucouvalas, 2002; Holzman & Pottenger, 2003; Ma, Prendinger, & Ishizuka, 2005; Mohammad, 2012; John, Boucouvalas, & Xu, 2006; Mihalcea & Liu, 2006; Genereux & Evans, 2006). Detecting these more subtle aspects of sentiment has wide-ranging applications, for example

in developing customer relation models, public health, military intelligence, and the video games industry, where it is necessary to make distinctions between anger and sadness (both of which are negative), calm and excited (both of which are positive), and so on.

3. Sentiment of Words

Term–sentiment associations have been captured by manually created sentiment lexicons as well as automatically generated ones.

3.1 Manually generated term-sentiment association lexicons

The General Inquirer (GI) has sentiment labels for about 3,600 terms (Stone, Dunphy, Smith, Ogilvie, & associates, 1966). Hu and Liu (2004) manually labeled about 6,800 words and used them for detecting sentiment of customer reviews. The MPQA Subjectivity Lexicon, which draws from the General Inquirer and other sources, has sentiment labels for about 8,000 words (Wilson, Wiebe, & Hoffmann, 2005). The NRC Emotion Lexicon has sentiment and emotion labels for about 14,000 words (Mohammad & Turney, 2010; Mohammad & Yang, 2011). These labels were compiled through Mechanical Turk annotations.[3]

For people, assigning a score indicating the degree of sentiment is not natural. Different people may assign different scores to the same target item, and it is hard for even the same annotator to remain consistent when annotating a large number of items. In contrast, it is easier for annotators to determine whether one word is more positive (or more negative) than the other. However, the latter requires a much larger number of annotations than the former (in the order of N^2, where N is the number of items to be annotated).

An annotation scheme that retains the comparative aspect of annotation while still requiring only a small number of annotations comes from survey analysis techniques and is called MaxDiff (Louviere, 1991). The annotator is presented with four terms and asked which word is the most positive and which is the least positive. By answering just these two questions five out of the six inequalities are known. If the respondent says that A is most positive and D is least positive, then:

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

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

3. https://www.mturk.com/mturk/welcome

will result in a list such that A and B are ranked close to each other and their real-valued association scores will also be close in value.

MaxDiff was used for obtaining annotations of relation similarity of pairs of items in a SemEval-2012 shared task (Jurgens, Mohammad, Turney, & Holyoak, 2012). Kiritchenko and Mohammad (2016a) applied Best–Worst Scaling to obtain real-valued sentiment association scores for words and phrases in three different domains: general English, English Twitter, and Arabic Twitter. They showed that on all three domains the ranking of words by sentiment remains remarkably consistent even when the annotation process is repeated with a different set of annotators. They also determine the minimum difference in sentiment association that is perceptible to native speakers of a language.

3.2 Automatically generated term-sentiment association lexicons

Semi-supervised and automatic methods have also been proposed to detect the polarity of words. Hatzivassiloglou and McKeown (1997) proposed an algorithm to determine the polarity of adjectives. SentiWordNet was created using supervised classifiers as well as manual annotation (Esuli & Sebastiani, 2006). Turney and Littman (2003) proposed a minimally supervised algorithm to calculate the polarity of a word by determining if its tendency to co-occur with a small set of positive seed words is greater than its tendency to co-occur with a small set of negative seed words. Mohammad, Kiritchenko, and Zhu (2013b) employed the Turney method to generate a lexicon (Hashtag Sentiment Lexicon) from tweets with certain sentiment-bearing seed-word hashtags such as (*#excellent, #good, #terrible,* and so on) and another lexicon (Hashtag Sentiment Lexicon) from tweets with emoticons.[4] Since the lexicons themselves are generated from tweets, they even have entries for the creatively spelled words (e.g. *happpeee*), slang (e.g. *bling*), abbreviations (e.g. *lol*), and even hashtags and conjoined words (e.g. *#loveumom*). Cambria, Olsher, and Rajagopal (2014) created SenticNet that has sentiment entries for 30,000 words and multi-word expressions using information propagation to connect various parts of common-sense knowledge representations. Kiritchenko et al. (2014b) proposed a method to create separate lexicons for words found in negated context and those found in affirmative context; the idea being that the same word contributes to sentiment differently depending on whether it is negated or not. These lexicons contain sentiment associations for hundreds of thousands of unigrams and bigrams. However, they do not explicitly handle combinations of terms with modals, degree adverbs, and intensifiers.

4. Sentiment of Phrases, Sentences, and Tweets: Sentiment Composition

Semantic composition, which aims at determining a representation of the meaning of two words through manipulations of their individual representations, has gained substantial attention in recent years with work from Mitchell and Loapata (2010), Baroni and Zamparelli (2010), Rudolph and Giesbrecht (2010), Yessenalina and Cardie (2011), Grefenstette, Dinu, Zhang, Sadrzadeh, and Baroni (2013), Grefenstette and Sadrzadeh (2011), and Turney (2014). Socher, Huval, Manning, and Ng (2012) and Mikolov, Sutskever, Chen, Corrado, and Dean (2013) introduced deep learning models and distributed word representations in

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

vector space (word embeddings) to obtain substantial improvements over the state-of-the-art in semantic composition. Mikolov's word2vec tool for generating word embeddings is available publicly.[5]

Sentiment of a phrase or a sentence is often not simply the sum of the sentiments of its constituents. Sentiment composition is the determining of sentiment of a multi-word linguistic unit, such as a phrase or a sentence, based on its constituents. Lexicons that include sentiment associations for phrases as well as their constituent words are referred to as *sentiment composition lexicons (SCLs)*. Kiritchenko and Mohammad created sentiment composition lexicons for English and Arabic that included: (1) negated expressions (Kiritchenko & Mohammad, 2016a, 2016b), (2) phrases with adverbs, modals, and intensifies (Kiritchenko & Mohammad, 2016a, 2016b), and (3) opposing polarity phrases (where at least one word in the phrase is positive and at least one word is negative, for example, *happy accident* and *dark chocolate*) (Kiritchenko & Mohammad, 2016c). Socher, Perelygin, Wu, Chuang, Manning, Ng, and Potts (2013) took a dataset of movie review sentences that were annotated for sentiment and further annotated ever word and phrasal constituent within those sentences for sentiment. Such datasets where sentences, phrases, and their constituent words are annotated for sentiment are helping foster further research on how sentiment is composed. We discuss specific types of sentiment composition, and challenges for automatic methods that address them, in the sub-sections below.

4.1 Negated Expressions

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 analysis of sentiment. Earlier works on negation handling employ simple heuristics such as flipping the polarity of the words in a negator's scope (Kennedy & Inkpen, 2005; Choi & Cardie, 2008) or changing the degree of sentiment of the modified word by a fixed constant (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Zhu, Guo, Mohammad, and Kiritchenko (2014) 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. 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).

Several aspects about negation are still not understood though: for example, can negators be ranked in terms of their average impact on the sentiment of their scopes (which negators impact sentiment more and which impact sentiment less); in what contexts does the same negator impact the sentiment of its scope more and in what contexts is the impact less; how do people in different communities and cultures use negations differently; and

5. https://code.google.com/p/word2vec

how negations of sentiment expressions should be dealt with by paraphrase and textual entailment systems.

4.2 Phrases with Degree Adverbs, Intensifiers, and 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. (See Kiritchenko and Mohammad (2016b) for some recent work exploring these questions in manually annotated sentiment composition lexicons.)

Modals (a kind of auxiliary verb) are used to convey the degree of confidence, permission, or obligation to the predicate. Thus, if the predicate is sentiment bearing, then the sentiment of the combination of the modal and the predicate can be different from the sentiment of the predicate alone. For example, *cannot work* seems less positive than *work* or *will work* (*cannot* and *will* are modals). There is little work on automatically determining the impact of modals on sentiment.

4.3 Sentiment of Sentences, Tweets, and SMS messages

Bag-of-word models such as the NRC-Canada system (Mohammad et al., 2013a; Kiritchenko, Zhu, Cherry, & Mohammad, 2014a; Kiritchenko et al., 2014b) and Unitn (Severyn & Moschitti, 2015) have been very successful in recent shared task competitions on determining sentiment of whole tweets, SMS messages, and sentences. However, approaches that apply systematic sentiment composition of smaller units to determine sentiment of sentences are growing in popularity. Socher et al. (2013) proposed a word-embeddings based model that learns the *sentiment* of term compositions. They obtain state-of-the-art results in determining both the overall sentiment and sentiment of constituent phrases in movie review sentences. This has inspired tremendous interest in more embeddings-based work for sentiment composition (Dong, Wei, Zhou, & Xu, 2014; Kalchbrenner, Grefenstette, & Blunsom, 2014). These recursive models do not require any hand-crafted features or semantic knowledge, such as a list of negation words or sentiment lexicons. However, they are computationally intensive and need substantial additional annotations (word and phrase-level sentiment labeling). Nonetheless, use of word-embeddings in sentiment composition

is still in its infancy, and we will likely see much more work using these techniques in the future.

4.4 Sentiment in Figurative Expressions

Figurative expressions in text, by definition, are not compositional. That is, their meaning cannot fully be derived from the meaning of their components in isolation. There is growing interest in detecting figurative language, especially irony and sarcasm (Carvalho, Sarmento, Silva, & De Oliveira, 2009; Reyes, Rosso, & Veale, 2013; Veale & Hao, 2010; Filatova, 2012; González-Ibánez, Muresan, & Wacholder, 2011). In 2015, a SemEval shared task was organized on detecting sentiment in tweets rich in metaphor and irony (Task 11).[6] 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 a large majority is negative; thereby suggesting that ironic tweets are largely negative. The SemEval 2014 shared task Sentiment Analysis in Twitter (Rosenthal et al., 2014a) 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. We found little to no work exploring automatic sentiment detection in hyperbole, understatement, rhetorical questions, and other creative uses of language.

5. Challenges in Annotating for Sentiment

Clear and simple instructions are crucial for obtaining high-quality annotations. This is true even for seemingly simple annotation tasks, such as sentiment annotation, where one is to label instances as positive, negative, or neutral. For word annotations, researchers have often framed the task as 'is this word positive, negative, or neutral?' (Hu & Liu, 2004), 'does this word have associations with positive, negative, or neutral sentiment?' (Mohammad & Turney, 2013), or 'which word is more positive?'/'which word has a greater association with positive sentiment' (Kiritchenko, Mohammad, & Salameh, 2016; Kiritchenko & Mohammad, 2016c). Similar instructions are also widely used for sentence-level sentiment annotations—'is this sentence positive, negative, or neutral?' (Rosenthal, Nakov, Kiritchenko, Mohammad, Ritter, & Stoyanov, 2015; Rosenthal, Ritter, Nakov, & Stoyanov, 2014b; Mohammad et al., 2016a; Mohammad, Zhu, Kiritchenko, & Martin, 2015b). We will refer to such annotation schemes as *the simple sentiment questionnaires*. On the one hand, this characterization of the task is simple, terse, and reliant on the intuitions of native speakers of a language (rather than biasing the annotators by providing definitions of what it means to be positive, negative, and neutral). On the other hand, the lack of specification leaves the annotator in doubt over how to label certain kinds of instances—for example, sentences where one side wins against another, sarcastic sentences, or retweets.

A different approach to sentiment annotation is to ask respondents to identify the target of opinion, and the sentiment towards this target of opinion (Pontiki, Papageorgiou, Galanis,

6. The proceedings will be released later in 2015.

Androutsopoulos, Pavlopoulos, & Manandhar, 2014b; Mohammad et al., 2015b; Deng & Wiebe, 2014). We will refer to such annotation schemes as *the semantic-role based sentiment questionnaires*. This approach of sentiment annotation is more specific, and more involved, than the simple sentiment questionnaire approach; however, it too is insufficient for handling several scenarios. Most notably, the emotional state of the speaker is not under the purview of this scheme. Many applications require that statements expressing positive or negative emotional state of the speaker should be marked as 'positive' or 'negative', respectively. Similarly, many applications require statements that describe positive or negative events or situations to be marked as 'positive' or 'negative', respectively. Instructions for annotating opinion towards targets do not specify how such instances are to be annotated, and worse still, possibly imply that such instances are to be labeled as neutral.

Some sentence types that are especially challenging for sentiment annotation (using either the simple sentiment questionnaire or the semantic-role based sentiment questionnaire) are listed below:

- *Speaker's emotional state:* The speaker's emotional state may or may not have the same polarity as the opinion expressed by the speaker. For example, a politician's tweet can imply both a negative opinion about a rival's past indiscretion, and a joyous mental state as the news will impact the rival adversely.

- *Success or failure of one side w.r.t. another*: Often sentences describe the success or failure of one side w.r.t. another side—for example, *'Yay! France beat Germany 3–1', 'Supreme court judges in favor of gay marriage'*, and *'the coalition captured the rebels'*. If one supports France, gay marriage, and the coalition, then these events are positive, but if one supports Germany, marriage as a union only between man and woman, and the rebels, then these events can be seen as negative.
  Also note that the framing of an event as the success of one party (or as the failure of another party) does not automatically imply that the speaker is expressing positive (or negative) opinion towards the mentioned party. For example, when Finland beat Russia in ice hockey in the 2014 Sochi Winter Olympics, the event was tweeted around the world predominantly as "Russia lost to Finland" as opposed to "Finland beat Russia". This is not because the speakers were expressing negative opinion towards the Russian team, but rather simply because Russia, being the host nation, was the focus of attention and traditionally Russian hockey teams have been strong.

- *Neutral reporting of valenced information*: If the speaker does not give any indication of her own emotional state but describes valenced events or situations, then it is unclear whether to consider these statements as neutral unemotional reporting of developments or whether to assume that the speaker is in a negative emotional state (sad, angry, etc.). Example:

  The war has created millions of refugees.

- *Sarcasm and ridicule*: Sarcasm and ridicule are tricky from the perspective of assigning a single label of sentiment because they can often indicate positive emotional state of the speaker (pleasure from mocking someone or something) even though they have a negative attitude towards someone or something.

- *Different sentiment towards different targets of opinion*: The speaker may express opinion about multiple targets, and sentiment towards the different targets might be different. The targets may be different people or objects (for example, an iPhone vs. an android phone), or they may be different aspects of the same entity (for example, quality of service vs. quality of food at a restaurant).

- *Precisely determining the target of opinion*: Sometimes it is difficult to precisely identify the target of opinion. For example, consider:

  Glad to see Hillary's lies being exposed.

  It is unclear whether the target of opinion is 'Hillary', 'Hillary's lies', or 'Hillary's lies being exposed'. One reasonable interpretation is that positive sentiment is expressed about 'Hillary's lies being exposed'. However, one can also infer that the speaker has a negative attitude towards 'Hillary's lies' and probably 'Hillary' in general. It is unclear whether annotators should be asked to provide all three opinion–target pairs or only one (in which case, which one?).

- *Supplications and requests*: Many tweets convey positive supplications to God or positive requests to people in the context of a (usually) negative situation. Examples include:

  May god help those displaced by war.
  Let us all come together and say no to fear mongering and divisive politics.

- *Rhetorical questions*: Rhetorical questions can be treated simply as queries (and thus neutral) or as utterances that give away the emotional state of the speaker. For example, consider:

  Why do we have to quibble every time?

  On the one hand, this tweet can be treated as a neutral question, but on the other hand, it can be seen as negative because the utterance betrays a sense of frustration on the part of the speaker.

- *Quoting somebody else or re-tweeting*: Quotes and retweets are difficult to annotate for sentiment because it is often unclear and not explicitly evident whether the one who quotes (or retweets) holds the same opinions as that expressed by the quotee.

The challenges listed above can be addressed to varying degrees by providing instructions to the annotators on how such instances are to be labeled. However, detailed and complicated instructions can be counter-productive as the annotators may not understand or may not have the inclination to understand the subtleties involved. See Mohammad (2016a) for annotation schemes that address some of these challenges.

6. Challenges in Multilingual Sentiment Analysis

Work on multilingual sentiment analysis has mainly addressed mapping sentiment resources from English into morphologically complex languages. Mihalcea, Banea, and Wiebe (2007)

use English resources to automatically generate a Romanian subjectivity lexicon using an English–Romanian dictionary. The generated lexicon is then used to classify Romanian text. Wan (2008) translated Chinese customer reviews to English using a machine translation system. The translated reviews are then annotated using rule-based system that uses English lexicons. A higher accuracy is achieved when using ensemble methods and combining knowledge from Chinese and English resources. Balahur and Turchi (2014) conducted a study to assess the performance of statistical sentiment analysis techniques on machine-translated texts. Opinion-bearing phrases from the New York Times Text (2002–2005) corpus were automatically translated using publicly available machine-translation engines (Google, Bing, and Moses). Then, the accuracy of a sentiment analysis system trained on original English texts was compared to the accuracy of the system trained on automatic translations to German, Spanish, and French. The authors conclude that the quality of machine translation is acceptable for sentiment analysis to be performed on automatically translated texts. Salameh, Mohammad, and Kiritchenko (2015) conducted experiments to determine loss in sentiment predictability when they translate Arabic social media posts into English, manually and automatically. As benchmarks, they use manually and automatically determined sentiment labels of the Arabic texts. They show that sentiment analysis of English translations of Arabic texts produces competitive results, w.r.t. Arabic sentiment analysis. They also claim that even though translation significantly reduces human ability to recover sentiment, automatic sentiment systems are affected relatively less by this.

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.

7. Challenges in Applying Sentiment Analysis

Applications of sentiment analysis benefit from the fact that even though systems are not extremely accurate at determining sentiment of individual sentences, they can accurately capture significant changes in the proportion of instances that are positive (or negative). It is also worth noting that such sentiment tracking systems are more effective when incorporating carefully chosen baselines. For example, knowing the percentage of tweets that are negative towards Russian President, Vladimir Putin, is less useful than, for instance, knowing: the percentage of tweets that are negative towards Putin before vs. after the invasion of Crimea; or, the percentage of tweets that are negative towards Putin in Russia vs. the rest of the world; or, the percentage of tweets negative towards Putin vs. Barack Obama (US president).

Sentiment analysis is commonly applied in several areas including tracking sentiment towards products, movies, politicians, and companies (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Pang & Lee, 2008), improving customer relation models (Bougie, Pieters, & Zeelenberg, 2003), detecting happiness and well-being (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, Agrawal, Park, et al., 2013), tracking the stock market (Bollen, Mao, & Zeng, 2011), and improving automatic dialogue systems (Velásquez, 1997; Ravaja, Saari, Turpeinen, Laarni, Salminen, & Kivikangas, 2006). The sheer volume of work in

this area precludes detailed summarization here. Nonetheless, it should be noted that often the desired application can help direct certain design choices in the sentiment analysis system. For example, the threshold between neutral and positive sentiment and the threshold between neutral and negative sentiment can be determined empirically by what is most suitable for the target application. Similarly, as suggested earlier, some applications may require only the identification of strongly positive and strongly negative instances.

Abundant availability of product reviews and their ratings has powered a lot of the initial research in sentiment analysis, however, as we look forward, one can be optimistic that the future holds more diverse and more compelling applications of sentiment analysis. Some recent examples include predicting heart attack rates through sentiment word usage in tweets (Eichstaedt, Schwartz, Kern, Park, Labarthe, Merchant, Jha, Agrawal, Dziurzynski, Sap, et al., 2015), corpus-based poetry generation (Colton, Goodwin, & Veale, 2012), generating music that captures the sentiment in novels (Davis & Mohammad, 2014), confirming theories in literary analysis (Hassan, Abu-Jbara, & Radev, 2012), and automatically detecting Cyber-bullying (Nahar, Unankard, Li, & Pang, 2012).

References

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

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

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.

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

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

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

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1193.

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.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

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.

Bougie, J. R. G., Pieters, R., & Zeelenberg, M. (2003). Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from tilburg university, Tilburg University.

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

Cambria, E., Olsher, D., & Rajagopal, D. (2014). Senticnet 3: a common and commonsense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.

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.

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

Colton, S., Goodwin, J., & Veale, T. (2012). Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, pp. 95–102.

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

Deng, L., & Wiebe, J. (2014). Sentiment propagation via implicature constraints.. In *EACL*, pp. 377–385.

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

Dong, L., Wei, F., Zhou, M., & Xu, K. (2014). Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis. In *Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI). AAAI*.

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*.

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.

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

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.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Comput. Linguist.*, *28*(3), 245–288.

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

Grefenstette, E., Dinu, G., Zhang, Y.-Z., Sadrzadeh, M., & Baroni, M. (2013). Multi-step regression learning for compositional distributional semantics. *arXiv preprint arXiv:1301.6939*.

Grefenstette, E., & Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1394–1404.

Hassan, A., Abu-Jbara, A., & Radev, D. (2012). Extracting signed social networks from text. In *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, pp. 6–14.

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.

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

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

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.

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*.

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)*.

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. (2016c). Sentiment composition of words with opposing polarities. 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., & Salameh, M. (2016). Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval-2016, San Diego, California.

Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. (2014a). Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442, Dublin, Ireland.

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

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.

Li, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, *41*(4, Part 2), 1742–1749.

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

Lloret, E., Balahur, A., Palomar, M., & Montoyo, A. Towards building a competitive opinion summarization system: challenges and keys. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: S pages=72–77, year=2009*.

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

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

Ma, C., Prendinger, H., & Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In Tao, J., & Picard, R. W. (Eds.), *First International*

Conference on Affective Computing and Intelligent Interaction (ACII-2005), pp. 622–628, Beijing, China.

Màrquez, L., Carreras, X., Litkowski, K. C., & Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, *34*(2), 145–159.

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

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

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

Mitchell, J., & Loapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388–1429.

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

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

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

Mohammad, S. M. (2012). #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. (2016a). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Mohammad, S. M. (2016b). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Meiselman, H. (Ed.), *Emotion Measurement*. Elsevier.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016a). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2016b). Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media, Submitted*.

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. (2013). Crowdsourcing a word-emotion association lexicon.. *29*(3), 436–465.

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

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015b). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, *51*(4), 480–499.

Nahar, V., Unankard, S., Li, X., & Pang, C. (2012). Sentiment analysis for effective detection of cyber bullying. In *Web Technologies and Applications*, pp. 767–774. Springer.

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

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

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.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

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

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.

Palmer, M., Gildea, D., & Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1–103.

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

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

Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., & Manandhar, S. (2014b). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

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

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.

Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., & Kivikangas, M. (2006). Spatial presence and emotions during video game playing: Does it matter with whom you play?. *Presence: Teleoperators and Virtual Environments*, *15*(4), 381–392.

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

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015). SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pp. 450–462, Denver, Colorado.

Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2014a). 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.

Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014b). SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 73–80, Dublin, Ireland.

Rudolph, S., & Giesbrecht, E. (2010). Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 907–916.

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

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

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

Severyn, A., & Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pp. 464–469.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '12, Jeju, Korea.

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.

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

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

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

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

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.

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*.

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

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

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

Turney, P. D. (2014). Semantic composition and decomposition: From recognition to generation. *arXiv preprint arXiv:1405.7908*.

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

Velásquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pp. 10–15. AAAI Press.

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

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

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.

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.

Yessenalina, A., & Cardie, C. (2011). Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 172–182.

Zarrella, G., & Marsh, A. (2016). MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

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. (2011). Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pp. 575–580.

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.

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

Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014). 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.