Information Processing and Management 51 (2015) 480–499

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Sentiment, emotion, purpose, and style in electoral tweets

Saif M. Mohammad *, Xiaodan Zhu, Svetlana Kiritchenko, Joel Martin

National Research Council Canada, Ottawa, Ontario K1A 0R6, Canada

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form 6 June 2014
Accepted 18 September 2014
Available online 16 October 2014

Keywords:
Sentiment analysis
Tweets
Elections
Emotions
Purpose
Crowdsourcing

ABSTRACT

Social media is playing a growing role in elections world-wide. Thus, automatically analyzing electoral tweets has applications in understanding how public sentiment is shaped, tracking public sentiment and polarization with respect to candidates and issues, understanding the impact of tweets from various entities, etc. Here, for the first time, we automatically annotate a set of 2012 US presidential election tweets for a number of attributes pertaining to sentiment, emotion, purpose, and style by crowdsourcing. Overall, more than 100,000 crowdsourced responses were obtained for 13 questions on emotions, style, and purpose. Additionally, we show through an analysis of these annotations that purpose, even though correlated with emotions, is significantly different. Finally, we describe how we developed automatic classifiers, using features from state-of-the-art sentiment analysis systems, to predict emotion and purpose labels, respectively, in new unseen tweets. These experiments establish baseline results for automatic systems on this new data.

1. Introduction

Elections, the cornerstone of democratic process, occur across the globe and often involve tens of millions of potential voters. Social media platforms such as Twitter give new opportunities to the electorate, the politicians, news corporations, and other participants to make their voice directly accessible to a large audience. However, the number of posts pertaining to a single event or topic such as a national election can grow to the hundreds of millions. The large number of tweets negates the possibility of a single person reading all of them to gain an overall global perspective. Thus, automatically analyzing electoral tweets, and specifically, analyzing sentiment and emotions in electoral tweets, can be beneficial for a number of downstream applications:

- *Understanding the role of target entities*: A number of entities tweet during elections, for example, the politicians, the voters, the disenfranchised, news corporations, non-governmental organizations, special interest groups, etc. Analyzing the extent to which tweets from various entities help shape public sentiment will improve our understanding of how social media is used during elections. It is also of interest to identify which portions of the voting electorate tweet about politics during elections. For example, some studies have shown that the more partisan electorate tend to tweet more, as do members from minority groups (Lassen & Brown, 2011).

* Corresponding author.
E-mail addresses: saif.mohammad@nrc-cnrc.gc.ca (S.M. Mohammad), xiaodan.zhu@nrc-cnrc.gc.ca (X. Zhu), svetlana.kiritchenko@nrc-cnrc.gc.ca (S. Kiritchenko), joel.martin@nrc-cnrc.gc.ca (J. Martin).

http://dx.doi.org/10.1016/j.ipm.2014.09.003
0306-4573/Crown Copyright © 2014 Published by Elsevier Ltd. All rights reserved.

- *Determining how public sentiment is shaped*: Some tweets (or some sets of tweets) have more impact in shaping public opinion than others. Determining characteristics of influential tweets is particularly useful.
- *Nowcasting and forecasting*: Tweet streams have been shown to help identify current public opinion towards the candidates in an election (nowcasting) (Conover, Goncalves, Ratkiewicz, Flammini, & Menczer, 2011; Golbeck & Hansen, 2011). Some research has also shown the predictive power of analyzing electoral tweets to determine the number of votes a candidate will get (forecasting) (Bermingham & Smeaton, 2011; Lampos, Preotiuc-Pietro, & Cohn, 2013; Tumasjan, Sprenger, Sandner, & Welpe, 2010a), however, other research expresses skepticism at the extent to which this is possible (Avello, 2012).
- *Identifying key electoral issues*: Electoral tweets can be analyzed to determine the extent to which voters are concerned about particular issues. For example, does the electorate value economic development much more than environment protection, and to what extent? Other related problems include identifying contentious issues (Maynard & Funk, 2011) and detecting voter polarization (Conover et al., 2011).
- *Impact of fake tweets*: Often during elections there is an increase of artificially generated tweets from twitterbots, botnets, and sock-puppets. Understanding the impact of these tweets on public sentiment and automatic methods to filter out such tweets are both important research problems.

Contributions of this work: Traditional information retrieval systems usually identify facts such as what a person is doing, at what time, in what location, etc. In this paper we analyze electoral tweets for more subtly expressed information such as sentiment (positive or negative), the emotion (joy, sadness, anger, etc.), the purpose or intent behind the tweet (to point out a mistake, to support, to ridicule, etc.), and the style of the tweet (simple statement, sarcasm, hyperbole, etc.). To our knowledge, this is the first tweets dataset annotated for all of these phenomena. We also developed two automatic statistical systems that use the annotated data for training and predict emotion and purpose labels in new unseen tweets. These experiments establish baseline results for automatic systems on this new data.

Data Annotation: We designed two detailed online questionnaires and annotated the tweets by crowdsourcing to Amazon's Mechanical Turk.[1] We obtained over 100,000 responses from about 3000 annotators. We present an extensive analysis of the annotations which lend support to interesting conclusions such as electoral tweets almost always express the emotion of the tweeter as opposed to somebody else's, the predominant emotion in these tweets is disgust followed by trust, electoral tweets convey negative emotions twice as often as positive emotions, and that different intents of tweeting may be associated with the same emotion. All the data created as part of this project: about 100,000 responses to questions about emotions, purpose, and style in electoral tweets are made available:

http://www.purl.org/net/PoliticalTweets2012.

Automatic Classifiers: We developed a classifier for emotion detection that obtains an accuracy of 56.84%. We show how the stimulus identification task can be framed as a classification task that circumvents more complicated problems of detecting entity mentions and coreferences. On this stimulus classification task, our supervised classifier obtains an F-score of 58.30.

We show that emotion detection alone can fail to distinguish between several different types of purpose. For example, the same emotion of disgust can be associated with many different kinds of purpose such as 'to criticize', 'to vent', and 'to ridicule'. Thus, detecting purpose provides information that is not obtained simply by detecting sentiment or emotion. We developed a preliminary system that automatically classifies electoral tweets as per their purpose, using various features that have traditionally been used in tweet classification, such as word ngrams and elongated words, as well as features pertaining to eight basic emotions. We show that resources developed for emotion detection are also helpful for detecting purpose. We then add to this system features pertaining to hundreds of fine emotion categories. We show that these features lead to significant improvements in accuracy above and beyond those obtained by the competitive preliminary system. The system obtains an accuracy of 44.58% on a 11-class task and an accuracy of 73.91% on a 3-class task. The various emotion lexicons are made freely available.[2]

The rest of the paper is organized as follows. In Section 2, we present related work. Section 3 presents the data annotation step and also a detailed analysis of the annotations obtained. In Section 4, we describe an automatic classifier for detecting emotions (Section 4.1), an experiment showing that emotion detection although related to purpose detection is in fact a different task (Section 4.2), and finally a classifier for detecting purpose (Section 4.3). Section 5 presents conclusions and directions for future work.

2. Related work

Related work is organized into two sub-sections: (1) on annotating text for sentiment, emotion, style, and categories such as purpose, and (2) on automatic classifiers for detecting these categories.

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

[2] http://www.purl.org/net/NRCEmotionLexicon.

2.1. Annotating text for sentiment, emotion, and style

In the past, separate efforts have been made to annotate text for sentiment, emotion, individual aspects of style such as sarcasm and irony, and categorization by goals.

Sentiment and Emotion: Wiebe and Cardie (2005) annotated sentences in newspaper articles for sentiment (positive, negative, or neutral). Pang and Lee (2008) (Section 7) present a detailed survey of sentiment annotation effort on product reviews and blog posts. Strapparava and Mihalcea (2007) annotated newspaper headlines for sentiment and emotion. They used six emotions that (Ekman, 1992) argued to be the most basic—joy, sadness, anger, fear, disgust, and surprise. Mohammad (2012b) used emotion-word hashtags such as #joy, #sadness, and #anger to compile tweets associated with eight basic emotions proposed by Plutchik (1980)—Ekman's six, trust, and anticipation. There has also been work on manually creating word–emotion association lexicons, for example, the NRC Emotion Lexicon (Mohammad & Turney, 2010), WordNet Affect (Strapparava & Valitutti, 2004), and the Affective Norms for English Words.[3]

Style: Filatova (2012) annotated Amazon product reviews for sarcasm and irony by crowdsourcing. Davidov, Tsur, and Rappoport (2010) collected tweets and Amazon product reviews and annotated them for sarcasm. González-Ibáñez, Muresan, and Wacholder (2011), Liebrecht, Kunneman, and van den Bosch (2013) used tweets with #sarcasm and #sarcastic as labeled data for sarcasm. Carvalho, Sarmento, Silva, and de Oliveira (2009) labeled newspaper sentences with irony. Reyes, Rosso, and Veale (2013) used #irony to compile a set of ironic tweets. We manually label tweets into one of eight categories of style including sarcasm, exaggeration, irony, and understatement. To our knowledge, this is the first such dataset.

Purpose-like Categories: There is no related work on detecting purpose from tweets. Below are some tweet annotation efforts for purpose-like categories. Naaman, Boase, and Lai (2010) organized 3379 tweets into the categories of information sharing, self promotion, opinions, statements, me now, questions, presence maintenance, and anecdote. Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010) annotated 5407 tweets into news, events, opinions, deals and private messages.

Here for the first time, we annotate the same dataset (electoral tweets) for a wide variety of sentiment, emotion, purpose, and style labels.

2.2. Classifiers for sentiment, emotion, and purpose-like categories

Over the last decade, there has been an explosion of work exploring sentiment analysis. Surveys by Pang and Lee (2008), Liu and Zhang (2012), Martínez-Cámara, Martín-Valdivia, Ureñalópez, and Montejoráez (2012) give summaries. (The Martinez survey focuses specifically on tweets.) Two state-of-the-art approaches particularly noteworthy and relevant to social media posts are: the NRC-Canada system (Kiritchenko, Zhu, & Mohammad, 2014; Mohammad, Kiritchenko, & Zhu, 2013a; Zhu, Kiritchenko, & Mohammad, 2014), which uses many lexical-features including those from sentiment lexicons, and the deep-learning based Stanford system (Socher, Huval, Manning, & Ng, 2012; Socher et al., 2013). Since emotion classifiers often benefit from the same kinds of features and system configurations as used in sentiment classifiers, we build our emotion classifier by drawing heavily from the NRC-Canada sentiment analysis system.

Sentiment: The NRC-Canada sentiment system was designed to detect the sentiment of short informal textual messages such as tweets. The system is based on a supervised statistical text classification approach leveraging a variety of surface-form, semantic, and sentiment features generated from word and character ngrams, manually created and automatically generated sentiment lexicons, parts of speech, word clusters, and Twitter-specific encodings such as hashtags and creatively spelled words and abbreviations (*yummeee, lol, etc.*). The sentiment features are primarily derived from novel high-coverage tweet-specific sentiment lexicons. These lexicons are automatically generated from tweets with sentiment-word hashtags and from tweets with emoticons. The system achieved the best results in SemEval-2013 shared task on Sentiment Analysis in Twitter (Task 2) (Wilson et al., 2013) and again in 2014 when the shared task was repeated (Task 9) (Rosenthal, Nakov, Ritter, & Stoyanov, 2014).

More complex approaches, such as recursive deep models (Socher et al., 2012, 2013) work in a bottom-top fashion over a parse-tree structure of a sentence to infer the sentiment label of the sentence as a composition of the sentiment expressed by its constituting parts: words and phrases. These models do not require any hand-crafted features or semantic knowledge, such as a list of negation words. However, they are computationally intensive and need substantial additional annotations (word and phrase-level sentiment labeling) to produce competitive results.

Emotion: Automatic emotion detection has been proposed for many different kinds of text (Aman & Szpakowicz, 2007; Boucouvalas, 2002; Genereux & Evans, 2006; Holzman & Pottenger, 2003; John, Boucouvalas, & Xu, 2006; Ma, Prendinger, & Ishizuka, 2005; Mihalcea & Liu, 2006; Mohammad, 2012a; Neviarouskaya, Prendinger, & Ishizuka, 2009; Tokuhisa, Inui, & Matsumoto, 2008; Zhe & Boucouvalas, 2002). More recently there has been work on tweets as well (Bollen, Pepe, & Mao, 2009; Choudhury, Counts, & Gamon, 2012; Kim, Gilbert, Edwards, & Graeff, 2009; Mohammad, 2012b; Tumasjan, Sprenger, Sandner, & Welpe, 2010b; Wang, Chen, Thirunarayan, & Sheth, 2012). Kim et al. (2009) analyzed sadness in Twitter posts reacting to news of Michael Jackson's death. Bollen et al. (2009) measured tension, depression, anger, vigor, fatigue, and confusion in tweets. Tumasjan et al. (2010b) study Twitter as a forum for political deliberation. Mohammad (2012b)

[3] http://csea.phhp.ufl.edu/media/anewmessage.html.

developed a classifier to identify emotions using tweets with emotion word hashtags as labeled data. However, none of this work explores the many semantic roles of emotion such as the who is feeling, what emotion, and towards whom.

Who, what sentiment/emotion, and towards whom/what: Detecting who is feeling, what emotion, and towards whom is essentially a semantic role-labeling problem (Gildea & Jurafsky, 2002). The semantic frame for 'emotions' in FrameNet (Baker, Fillmore, & Lowe, 1998) is shown in Table 1. (FrameNet has a database of more than 1000 semantic frames.) In this work, we focus on the roles of *Experiencer, State,* and *Stimulus*. Note that the state or emotion is often not explicitly present in text. Nonetheless, it is usually easy to deduce. Other roles such as *Reason, Degree,* and *Event* are also of significance, and remain suitable avenues for future work.

To the best of our knowledge, there exists no work on semantic role labeling of emotions in tweets. However, there is work on the related task of extracting opinions and the topics of opinions. Much of this work is focused on opinions about product features (Kessler & Nicolov, 2009 Popescu & Etzioni, 2005; Qadir, 2009; Su, Xiang, Wang, Sun, & Yu, 2006; Xu, Huang, & Wang, 2013; Zhang & Liu, 2011; Zhang, Liu, Lim, & O'Brien-Strain, 2010). In the 2014 SemEval shared task Aspect Based Sentiment Analysis (Task 4), automatic systems had to detect sentiment towards aspects terms in the Restaurant and Laptop domains.

Consider:

The lasagna was great, but the service left a lot to be desired.

We can gather from this sentence that the customer has a positive sentiment towards the lasagna, but a negative sentiment towards the service. With over 40 teams participating, the NRC-Canada system (Kiritchenko, Zhu, Cherry, & Mohammad, 2014) obtained the best results overall. We adapt this system to detect the stimulus of emotions in the electoral tweets data.

Purpose: To the best of our knowledge, there is no work yet on classifying electoral or political tweets by purpose. Collier, Son, and Nguyen (2011) classified flu-related tweets into avoidance behavior, increased sanitation, seeking pharmaceutical intervention, wearing a mask, and self reported diagnosis. Caragea et al. (2011) classified earthquake-related tweets into medical emergency, people trapped, food shortage, water shortage, water sanitation, shelter needed, collapsed structure, food distribution, hospital/clinic services, and person news. There exists work on determining political alignment of tweeters (Conover, Goncalves, et al., 2011; Golbeck and Hansen, 2011), identifying contentious issues and political opinions (Maynard & Funk, 2011), detecting the amount of polarization in the electorate (Conover, Ratkiewicz, et al., 2011), and detecting sentiment in political tweets (Bermingham & Smeaton, 2011; Chung & Mustafaraj, 2011; O'Connor, Balasubramanyan, Routledge, & Smith, 2010).

Other electoral tweets classifiers for downstream applications: Automatic analysis, especially of sentiment, emotion, and purpose, has applications such as determining political alignment of tweeters (Conover, Goncalves, et al., 2011; Golbeck and Hansen, 2011), identifying contentious issues and political opinions (Maynard & Funk, 2011), detecting the amount of polarization in the electorate (Conover, Ratkiewicz, et al., 2011), and even predicting the voting intentions or outcome of elections (Bermingham & Smeaton, 2011; Lampos et al., 2013; Tumasjan et al., 2010a). One of the more recent works (Lampos et al., 2013) analyzes tweets from UK and Austria and successfully predicts voting intention in more than 300 polls across the two countries. They filter out irrelevant tweets using a bilinear model on both the word space and the user space.

In this paper, our goal is to annotate a single electoral tweets dataset with a number of sentiment-, emotion-, and purpose-related labels, and establish baselines on what certain automatic classifiers can achieve when made to automatically identify these labels in previously unseen test data. We hope that the dataset and the classifiers will be useful in many of the downstream applications such as those listed above and earlier in the Introduction.

Table 1
The FrameNet frame for emotions. The roles examined in this paper are in bold.

Role	Description
Core:	
Event	The Event is the occasion or happening that Experiencers in a certain emotional state participate in
Experiencer	The Experiencer is the person or sentient entity that experiences or feels the emotions
Expressor	The body part, gesture, or other expression of the Experiencer that reflects his or her emotional state
State	The State is the abstract noun that describes a more lasting experience by the Experiencer
Stimulus	The Stimulus is the person, event, or state of affairs that evokes the emotional response in the Experiencer
Topic	The Topic is the general area in which the emotion occurs It indicates a range of possible Stimulus
Non-Core:	
Circumstances	The Circumstances is the condition(s) under which the Stimulus evokes its 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
Parameter	The Parameter is a domain in which the Experiencer experiences the Stimulus
Reason	The Reason is the explanation for why the Stimulus evokes a certain emotional response

Table 2
Query terms used to collect 2012 US presidential election tweets.

#4moreyears	#Barack	#campaign2012
#dems2012	#democrats	#election
#election2012	#gop2012	#gop
#joebiden2012	#mitt2012	#Obama
#ObamaBiden2012	#PaulRyan2012	#president
#president2012	#Romney	#republicans
#RomneyRyan2012	#veep2012	#VP2012
Barack	Obama	Romney

3. Data collection and annotation for sentiment, emotions, purpose, and style

In the subsections below we describe how we collected tweets pertaining to the 2012 US presidential elections and annotated them for sentiment, emotions, purpose, and style.

3.1. Identifying electoral tweets

We created a corpus of tweets by polling the Twitter Search API, during August and September 2012, for tweets that contained commonly known hashtags pertaining to the 2012 US presidential elections. Table 2 shows the query terms we used. Apart from 21 hashtags, we also collected tweets with the words Obama, Barack, or Romney. We used these additional terms because they are names of the two presidential candidates, and the probability that these words were used to refer to somebody else in tweets posted in August and September of 2012 was low.

The Twitter Search API was polled every four hours to obtain new tweets that matched the query. Close to one million tweets were collected, which we make freely available to the research community. The query terms which produced the highest number of tweets were those involving the names of the presidential candidates, as well as #election2012, #campaign, #gop, and #president.

We used the metadata tag "iso_language_code" to identify English tweets. Since this tag is not always accurate, we also discarded tweets that did not have at least two valid English words. We used the Roget Thesaurus as the English word inventory. This step also helps discard very short tweets and tweets with a large proportion of misspelled words. We discarded retweets, which can easily be identified through the presence of RT, rt, or Rt in the tweet (usually in the beginning of the post). Finally, there remained close to 170,000 original English tweets.

3.2. Annotating emotions through crowdsourcing

We used Amazon's Mechanical Turk and CrowdFlower to crowdsource the annotation of the electoral tweets.[4] The questionnaires posted on Mechanical Turk are called *HITs (human intelligence tasks)*. CrowdFlower acts as intermediary between job requesters and crowdsourcing platforms such as Mechanical Turk and SamaSource. It provides additional quality control measures such as a mechanism to avoid annotations from annotators who consistently fail to correctly respond to certain gold questions—questions for which the job requester has provided answers. Crowdflower recommends that the job requester provide gold answers to questions in 5–10% of the HITs.

We randomly selected about 2000 tweets, each by a different Twitter user. We set up two questionnaires on Mechanical Turk through CrowdFlower. The questions to be included in the two questionnaires as well as the options to be provided within each question were established after soliciting responses from our colleagues at the National Research Council Canada. We restricted annotations to Turkers from the United States. We also requested that only native speakers of English attempt the HITs.

The first questionnaire was used to determine the presence of emotions in a tweet, the style of the tweet, and the purpose of the tweet. It also had a question to verify whether the tweet was truly relevant to US politics. Below is an example:

Questionnaire 1: Emotions in the US election tweets
Tweet: Mitt Romney is arrogant as hell.

Q1. Which of the following best describes the **Emotions** in this tweet?
- This tweet expresses or suggests an emotional attitude or response to something.
- This tweet expresses or suggests two or more contrasting emotional attitudes or responses.
- This tweet has no emotional content.
- There is some emotion here, but the tweet does not give enough context to determine which emotion it is.
- It is not possible to decide which of the above options is appropriate.

[4] https://crowdflower.com.

Q2. Which of the following best describes the **Style** of this tweet?
- simple statement or question
- exaggeration or hyperbole
- sarcasm
- rhetorical question
- understatement
- weird, surreal, or off-the-wall
- humorous, but none of the above
- none of the above

Q3. Which of the following best describes the **Purpose** of this tweet?
- to point out hypocrisy or inconsistency
- to point out mistake or blunder
- to disagree
- to ridicule
- to criticize, but none of the above
- to vent
- to agree
- to praise, admire, or appreciate
- to support
- to motivate or to incite action
- to be entertaining
- to provide information without emotion
- none of the above

Q4. Is this tweet about US politics and elections?
- Yes, this tweet is about US politics and elections.
- No, this tweet has nothing to do with US politics or anybody involved in it.

We posted 2042 HITs corresponding to 2042 tweets. We requested responses from at least three annotators for each HIT. The response to a HIT by an annotator is called an *assignment*. In Mechanical Turk, an annotator may provide assignments for as many HITs as they wish. Thus, even though only three annotations are requested per HIT, about 400 annotators contributed assignments for the 2,042 tweets.

Observe that we implicitly grouped the options for Q3 into three coarse categories by putting extra vertical space between the groups. These coarse categories correspond to *oppose* (to point out hypocrisy, to point out mistake, to disagree, to ridicule, to criticize, to vent), *favor* (to agree, to praise, to support), and *other*. Even though there is some redundancy among the fine categories, they are more precise and may help annotation. Eventually, however, it may be beneficial to combine two or more categories for the purposes of automatic classification. The amount of combining will depend on the task at hand, and can be done to the extent that anywhere from eleven to two categories remain.

The tweets that were marked as having one emotion were chosen for annotation by Questionnaire 2. Here we asked various questions pertaining to emotional state such as who is feeling the emotion, what emotion, towards whom, and more. We decided to classify the emotions in tweets into one of eight basic emotion categories proposed by Plutchik. However, these categories are fairly coarse. For example, disgust, dislike, hate, disappointment, and indifference all fall under the category of disgust. In order to make annotation easier, we presented the Turkers with a larger list of 19 possible emotions. The annotations for 19 emotions were eventually mapped into the 8 basic emotions as follows: trust, acceptance, admiration, and like were mapped to trust; fear mapped to fear; surprise, uncertainty, amazement to surprise; sadness mapped to sadness; disgust, dislike, hate, disappointment, and indifference mapped to disgust; anger to anger; anticipation and vigilance to anticipation; joy and calmness to joy. The data analysis and automatic emotion classification we present in the following sections all work at the level of the eight basic emotions. Below is an example of Questionnaire 2:

Questionnaire 2: Who is feeling what, and towards whom?
Tweet: Mitt Romney is arrogant as hell.

Q1. Who is feeling or who felt an emotion?

Q2. What emotion? Choose one of the options from below that best represents the emotion.
- acceptance
- admiration or like
- amazement
- anger, hostility or fury
- anticipation or interest
- calmness or serenity
- disappointment
- disgust

- dislike
- fear, panic or terror
- hate
- indifference
- joy, happiness or elation
- like
- sadness or grief or sorrow
- surprise
- trust
- uncertainty or indecision
- vigilance

Q3. If there is a better word for describing the emotion (than the ones listed above), then type it here:

Q4. If when answering Q2 you have chosen an emotion from the "Other emotions" category or if you answered Q4, then please tell us if the emotion in this tweet is positive, negative, or neither?

- positive emotion
- negative emotion
- neither positive nor negative

Q5. How strongly is the emotion being expressed in this tweet?

- the emotion is being expressed with a high intensity
- the emotion is being expressed with medium intensity
- the emotion is being expressed with a low intensity

Q6. Towards whom or what? In other words, who or what is the stimulus of the emotion?

Q7. Which words in the tweet help in identifying the emotion?

Q8. What reason can be deduced from the tweet for the emotion? What is the cause of the emotion?

Q9. This tweet is about which of the following issues:

ECONOMY

- federal debt
- jobs
- housing
- taxes
- military spending
- About the Economy: but not related to any of the above issues.

CONFLICTS AND TERRORISM

- terrorism
- Afghanistan or Iraq war
- Arab Spring, Egypt, Syria, or Libya
- Iran, Israel, or Palestine
- About Conflicts and Terrorism: but not related to any of the above issues.

SOCIAL AND CIVIL ISSUES

- education
- environment
- gay rights
- gun control/rights
- health care
- racism
- religion
- women's rights
- About Social and Civil Issues: but not related to any of the above issues.

OTHER

- about the election process, election publicity, or election campaign
- none of the above.

Table 3
Questionnaire 1: Percentage of tweets in each category of Q1. Only those tweets that were annotated by at least two annotators were included. A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 1889 such tweets in total. The percentage of tweets in the largest category is shown in bold.

	Percentage of tweets
Suggests an emotional attitude	**87.98**
Suggests two contrasting attitudes	2.22
No emotional content	8.21
Some emotion; not enough context	1.32
Unknown; not enough context	0.26
All	100.0

Q10. If the tweet is about an issue not listed above, then type it here:

We requested responses from at least five annotators for each of these HITs.

After performing a small pilot annotation effort, we realized that the stimulus in most of the electoral tweets was one among a handful of entities. Thus we reformulated question 6 as shown below:

Q6b. Which of these best describes the target of the emotion?

- Barack Obama and/or Joe Biden
- Mitt Romney and/or Paul Ryan
- Some other individual
- Democratic party, democrats, or DNC
- Republican party, republicans, or RNC
- Some other institution
- Election campaign, election process, or elections
- The target is not specified in the tweet
- None of the above.

Even though it is possible that more than one option may apply for a tweet, we allowed the Turkers to select only one option for each question. We did this to encourage annotators to select the option that best answers the questions. We wanted to avoid situations where an annotator selects multiple options just because they are vaguely relevant to the question. Before going live, the survey was approved by the ethics committee at the National Research Council Canada. Both questionnaires, in exactly the form the Turkers saw it, are made publicly available.[5]

3.3. Annotation analyses

Apart from the quality control measures employed by CrowdFlower (the use of gold questions), we deployed additional measures to discard poor annotations. For each annotator and for each question, we calculated the probability with which the annotator agreed with the response chosen by the majority of the annotators. We identified poor annotators as those that had an agreement probability that was more than two standard deviations away from the mean. All annotations by these annotators were discarded.[6]

We determine whether a tweet is to be assigned a particular category based on strong majority. That is, a tweet belongs to category X only if more than half of the annotators agree with each other. A minimum of at least two annotations per tweet are required for Questionnaire 1, whereas a minimum of at least three annotations per tweet are required for Questionnaire 2.

Percentage of tweets in each of the five categories of Q1 is shown in Table 3. Observe that the majority category for Q1 is 'suggests an emotion'—87.98% of the tweets were identified as having an emotional attitude.

Table 4 gives the distributions of the various options for question 2 (style). Simple statements are predominant in electoral tweets, but a fair percentage of these tweets correspond to exaggeration and sarcasm. None of the tweets was marked as being an understatement.

Analysis of responses to Q3 revealed that the category 'to motivate or to incite action' was often confused with the category 'to support'. Thus we merged the two categories into one. Also, the category 'to be entertaining' was highly confused with many other options, thus we ignored this category completely. Percentage of tweets in eleven categories of Q3 is shown in Table 5. Observe that the majority category for purpose is 'to support'—26.49%. Fig. 1 gives the distributions of the three

[5] https://www.dropbox.com/sh/2zm58tm58erfkv1/AABhk5xa4wRPs3hN8ZYF-ky-a.

[6] A histogram of the annotations from Questionnaire 1 is shown in the Appendix A.

Table 4
Questionnaire 1: Percentage of tweets in each category of Q2 (Style). Only those tweet that were annotated by at least two annotators were included. A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 1569 such tweets in total. The percentage of tweets in the largest category is shown in bold.

	Percentage of tweets
Simple statement or question	**76.86**
Exaggeration or hyperbole	9.75
Sarcasm	7.39
Rhetorical question	3.19
Understatement	0.00
Weird, surreal, or off-the-wall	1.02
Humorous, but none of the above	1.66
None of the above	0.13
All	100.0

Table 5
Questionnaire 1: Percentage of tweets in each of the eleven categories of Q3 (Purpose). A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 1072 such tweets. Category "to support" includes "to motivate or to incite action". The percentage of tweets in the largest category is shown in bold.

Purpose of tweet	Percentage of tweets
Favor	
To agree	0.47
To praise, admire, or appreciate	15.02
To support	**26.49**
Oppose	
To point out hypocrisy or inconsistency	7.00
To point out mistake or blunder	3.45
To disagree	2.52
To ridicule	15.39
To criticize, but none of the above	7.09
To vent	8.21
Other	
To provide information without any emotional content	13.34
None of the above	1.03
All	100.0

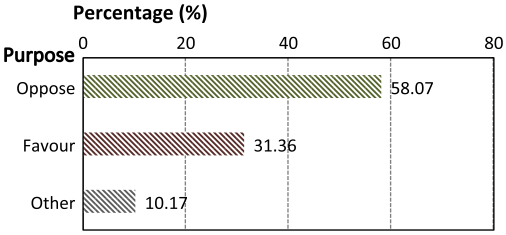

Fig. 1. Questionnaire 1: Percentage of tweets in each of the three coarse categories of Q3 (Purpose). A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 1672 such tweets. The total number of tweets is larger since the annotator agreement on the three categories is larger than on eleven categories.

coarse categories of purpose. Observe, that the political tweets express opposition (58.07%) much more often than favor (31.76%).

Responses to question 4, revealed that a large majority (95.56%) of the tweets are relevant to US politics and elections. Thus the hashtags shown earlier in Table 2 were effective in identifying political tweets.

As mentioned earlier, only those tweets that were marked as having an emotion (with high agreement) were annotated further through Questionnaire 2.[7] Responses to Q1 of Questionnaire 2, showed that in the vast majority of the instances (99.825%), the tweets contain emotions of the tweeter. The data did include some tweets that referred to emotions of others (Romney, GOP, and president), but these instances were rare.

[7] A histogram of the annotations for Questionnaire 2 is shown in the Appendix A.

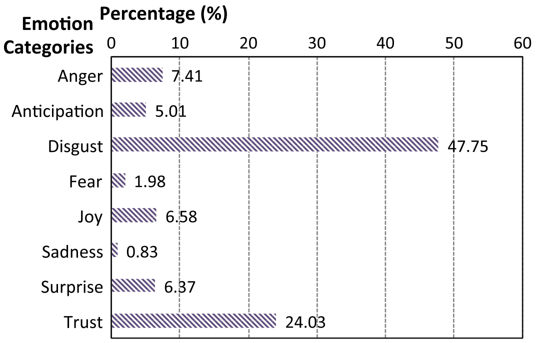

Fig. 2. Questionnaire 2: Percentage of tweets in each of the categories of Q2 (Emotion Categories). Only those tweets that were annotated by at least three annotators were included. A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 965 such tweets in total.

Figs. 2 and 3 give the distributions of the various options for Questions 2 and 6b of Questionnaire 2. Disgust (47.75%) is by far the most dominant emotion in the tweets of 2012 US presidential elections. The next most prominent emotion is that of trust (24.03%). About 58% of the tweets convey negative emotions towards someone or something. Fig. 3 shows that the stimulus of emotions was often one of the two presidential candidates (close to 55% of the time)—Obama: 29.90%, Romney: 24.87%.

Figs. 4 and 5 give the distributions of the options for Questions 4 and 5, respectively. Observe that most of the emotion expressions are of medium intensity (65.51% of the tweets). Question 5 was optional, and annotators were asked to respond only if the emotion chosen in question 2 is not one of the given positive or negative emotions. Fig. 5 shows that negative emotions were more dominant than the positive ones in the electoral tweets.

Table 6 gives the distributions of the options for Question 9. Observe that the distribution is highly skewed towards the class "election process" (76.08%). It is interesting to note that most of the tweets are not about key electoral issues such as

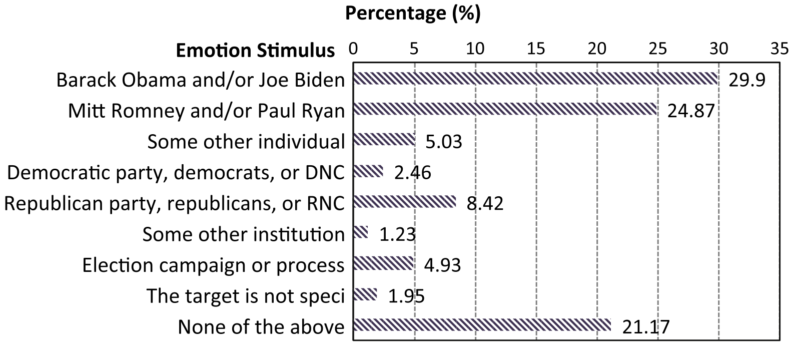

Fig. 3. Questionnaire 2: Percentage of tweets in the categories of Q6b (Emotion Stimulus). There were 973 tweets with strong majority.

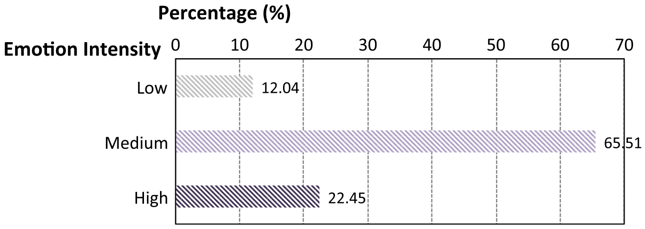

Fig. 4. Questionnaire 2: Percentage of annotations pertaining to the categories in Q4 (Emotion Intensity). A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 980 such tweets.

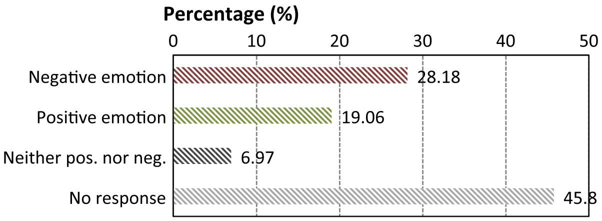

Fig. 5. Questionnaire 2: Percentage of annotations pertaining to the categories in Q5. A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 976 such tweets.

Table 6
Questionnaire 2: Percentage of annotations pertaining to the categories in Q9 (Issues). There were 974 tweets with strong majority. The percentage of tweets in the largest category is shown in bold.

	Percentage of tweets
Economy	
Federal debt	0.82
Jobs	1.85
Housing	0.00
Taxes	1.13
Military spending	0.21
About the Economy but not related to any of the above	2.05
Sub total	6.06
Conflicts and terrorism	
Terrorism	0.62
Afghanistan or Iraq war	0.10
Arab Spring, Egypt, Syria, or Libya	0.82
Iran, Israel, or Palestine	0.41
About Conflicts and Terrorism but not related to any of the above	0.82
Sub total	2.77
Social and civil issues	
Education	0.92
Environment	0.00
Gay rights	0.62
Gun control/rights	0.41
Health care	0.41
Racism	0.72
Religion	0.62
Women's rights	2.36
About social and civil issues but not related to any of the above	0.51
Sub total	6.57
Other	
About the election process, election publicity, or election campaign	**76.08**
None of the above	8.52
Total	100.00

terrorism, economy, and women's rights, but rather about the election process itself. Questions 3, 6, 7, 8, and 10 had free-text responses, and so we do not show their distributions.

3.3.1. Inter-annotator agreement

We calculated agreement statistics on the full set of annotations, and not just on the annotations with a strong majority as described in the previous section. Table 7 shows *inter-annotator agreement (IAA)* for the questions—the average percentage of times two annotators agree with each other. Another way to gauge agreement is by calculating the average probability with which an annotator picks the majority class. The last column in Table 7 shows the average probability of picking the majority class (APMS) by the annotators (higher numbers indicate higher agreement). Observe that there is high agreement on determining whether a tweet has an emotion or not, and on determining whether the tweet is related to the 2012 US presidential elections or not. The questions in Questionnaire 2 pertaining to the emotional state, stimulus, etc. were less straightforward and tend to require more context than just the target tweet for a clear determination, but yet the annotations had moderate agreement.

Table 7
Agreement statistics: inter-annotator agreement (IAA) and average probability of choosing the majority class (APMS).

	IAA	APMS
Questionnaire 1:		
Q1	78.02	0.845
Q2	55.77	0.688
Q3	43.58	0.520
Q4	96.76	0.974
Questionnaire 2:		
Q1	52.95	0.731
Q2	59.59	0.736
Q4	73.82	0.718
Q5	45.17	0.662
Q6b	44.47	0.641
Q9	69.91	0.819

4. Automatically detecting emotions and purpose

We now present automatic classifiers that use some of the annotations described above as training data and predict emotions and purpose in unseen test tweets. In the subsections below we present: (1) a basic automatic system to determine who is feeling what emotion, and towards whom (Section 4.1) (2) the correlations and distinctions between emotions and purpose (Section 4.2), (3) a basic automatic system to automatically classify tweets into eleven categories of purpose (Section 4.3). The objective of these experiments is to establish baseline results on this new electoral tweets dataset using features from state-of-the-art sentiment analysis systems, and also to establish the relationship between purpose and emotions in electoral tweets. We leave the automatic determination of style as well as other aspects of affect such as determining the reason behind the emotion for future work. The data and annotations are made freely available.

4.1. Automatically detecting semantic roles of emotions in tweets

Since in most instances (99.83%) the experiencer of emotions in a tweet is the tweeter, in this section we focus on automatically detecting the other two semantic roles: the emotional state and the stimulus.

We treat the detection of emotional state and stimulus as two subtasks for which we train state-of-the-art support vector machine (SVM) classifiers. SVM is a learning algorithm proved to be effective on many classification tasks and robust on large feature spaces. In our experiments, we exploited several different classifiers and found that SVM outperforms others such as maximum-entropy models (i.e., logistic regression). We also tested the most popular kernels such as the polynomial and RBF kernels with different parameters in 10-fold cross validation. We found that a simple linear kernel yielded the best performance. We used the LibSVM package (Chang & Lin, 2011a).

Our system builds on the classifiers and features used in two previous systems: (1) the system described in Mohammad (2012b) which was shown to perform significantly better at emotion detection than some other previous systems on the news paper headlines corpus and (2) the system described in (Mohammad, Kiritchenko, & Zhu, 2013b) which ranked first (among 44 participating teams) in a 2013 SemEval competition on detecting sentiment in tweets (Task 2) (Wilson et al., 2013). In each experiment below, we report results of ten-fold stratified cross-validation.

4.1.1. Detecting emotional state

Below we present the features used for detecting emotional state in electoral tweets and the results obtained with them.

Features: We included the following features for emotional state in tweets, where emotional state can be one of eight possible basic emotions: joy, sadness, anger, fear, surprise, anticipation, trust, and disgust.

- *Word ngrams*: unigrams (single words) and bigrams (two-word sequences). All words were stemmed with Porter's stemmer (Porter, 1980).
- *Punctuations*: number of contiguous sequences of exclamation marks, question marks, or a combination of them.
- *Elongated words*: the number of words with the ending character repeated more than 3 times, e.g., "soooo" and "mannnnnn". Elongated words have been used similarly by Brody and Diakopoulos (2011).
- *Emoticons*: presence/absence of positive and negative emoticons. The emoticon and its polarity were determined through a simple regular expression adopted from Christopher Potts' tokenizing script.[8] Emoticons have been used widely in the past for sentiment analysis applications by Go, Bhayani, and Huang (2009) and Mohammad et al. (2013a).

[8] http://sentiment.christopherpotts.net/tokenizing.html.

Table 8
Emotion detection task: overall results on the 8-class task.

	Accuracy
Random baseline	30.26
Majority baseline	47.75
Automatic SVM system	56.84
Human performance	69.80

- *Emotion Lexicons*: We used the NRC word–emotion association lexicon (Mohammad & Turney, 2010) to check if a tweet contains emotional words. The lexicon contains human annotations of emotion associations for about 14,200 word types. The annotation includes whether a word is positive or negative (sentiments), and whether it is associated with the eight basic emotions (joy, sadness, anger, fear, surprise, anticipation, trust, and disgust). If a tweet has three words that have associations with emotion joy, then the *LexEmo_emo_joy* feature takes a value of 3. The NRC lexicon was used for emotion classification of newspaper headlines by Mohammad (2012a), and such emotion lexicon features have been generated from WordNet Affect in earlier works (Alm & Ovesdotter, 2008; Aman & Szpakowicz, 2007).
- *Negation features*: We examined tweets to determine whether they contained negators such as *no, not,* and *should n't*. An additional feature determined whether the negator was located close to an emotion word (as determined by the emotion lexicon) in the tweet and in the dependency parse of the tweet. The list of negation words was adopted from Christopher Potts' sentiment tutorial.[9] Negation features of this kind were used earlier by Polanyi and Zaenen (2004), Kennedy and Inkpen (2005), Choi and Cardie (2008), Taboada, Brooke, Tofiloski, Voll, and Stede (2011).
- *Position features*: We included a set of position features to capture whether the feature terms described above appeared at the beginning or the end of the tweet. For example, if one of the first five terms in a tweet is a joy word, then the feature *LexEmo_joy_begin* was triggered.
- *Combined features* Though non-linear models like SVM (with non-linear kernels) can capture interactions between features, we explicitly combined some of our features. For example, we concatenated all emotion categories found in a given tweet. If the tweet contained both surprise and disgust words, a binary feature *LexEmo_surprise_disgust* was triggered.

Results: Table 8 shows the results. We use accuracy as the evaluation metric. Note that accuracy equals micro-averaged P, R, and F in this case as the categories are mutually exclusive. We included two baselines: the random baseline corresponds to a system that randomly guesses the emotion of a tweet, whereas the majority baseline assigns all tweets to the majority category (disgust). Since the data is significantly skewed towards disgust, the majority baseline is relatively high.

The automatic system obtained an accuracy of 56.84%, which is significantly higher than the majority baseline. It should be noted that the highest scores in the SemEval 2013 task for detecting sentiment of tweets (Task 2) was around 69% (Mohammad et al., 2013b; Wilson et al., 2013). That task even though related involved only three classes (positive, negative, and neutral). Thus it is not surprising that for an 8-way classification task, the performance is somewhat lower. Further analysis showed that our system obtains highest accuracies on the categories of disgust (71.86%) and trust (54.51%). These were the two most frequent categories in the data (see Fig. 2), and thus the result is not surprising. Confusion matrices (not shown here) revealed that the positive emotions (joy and trust) tended to be confused for each other and many of tweets with non-disgust negative emotions were often marked as disgust.

As mentioned earlier, human annotators do not always agree with each other. To estimate the human performance, for each tweet we randomly sample a human annotation from its multiple annotations. We compare it with the majority category chosen from the remaining human annotations for that tweet. Such sampling is conducted over all tweets and then evaluated. The resulting accuracy is 69.80%.

Table 9 shows the results of ablation experiments—the accuracies obtained with one of the feature groups removed. The higher the drop in performance, the more useful is that feature. Observe that the ngrams are the most useful features, followed by the emotion lexicons. Most of the gain is due to word ngrams, but character ngrams provide small additional gains as well.[10] The NRC emotion lexicon improved results as well. Paying attention to negation was also beneficial, however, emotional encodings such as elongated words, emoticons, and punctuations did not help much. It is possible that much of the discriminating information they might have is already provided by unigram and character ngram features.

4.1.2. Detecting emotion stimulus

As discussed earlier, instead of detecting and labeling the original text spans, we ground the emotion stimulus directly to the predefined entities. This allows us to circumvent mention detection and co-reference resolution on linguistically less well-formed text. We treat the problem as a classification task, in which we classify a tweet into one of the eight categories defined in Fig. 3. The categories of 'target not specified' and 'none of the above' were combined to form the negative class.

[9] http://sentiment.christopherpotts.net/lingstruc.html.

[10] Note that word ngrams and character ngrams are redundant to some extent, which is why the drop in performance when removing them individually does not sum up to the drop in performance when removing ngrams as a whole.

Table 9
Emotion detection task: the accuracies obtained with one of the feature groups removed.

Experiment	Accuracy
All features	56.84
All – ngrams	53.35
All – word ngrams	54.44
All – character ngrams	56.32
All – emotion lexicon	54.34
All – negation	55.80
All – encodings	56.82
(encodings = elongated words, emoticons, punctuations, uppercase)	

Features: We used the features listed below for detecting emotion stimulus:

- *Word ngrams*: Same as described earlier for emotional state.
- *Lexical features*: We collected lexicons that contain a variety of words and phrases describing the categories in Fig. 3. For example, the Republican party may be called as "gop" or "Grand Old Party"; all such words or phrases are put into the lexicon called "republican". We counted how many words in a given tweet are from each of these lexicons.
- *Hashtag features*: Hashtags related to the U.S. election were collected. We organized them into different categories and use them to further smooth the sparseness. For example, *#4moreyear* and *#obama* are put into the same hashtag lexicon and any occurrence of such hashtags in a tweet triggers the feature *hashtag_obama_generalized*, indicating that this is a general version of hashtag related to president Barack Obama.
- *Position features*: Same as described earlier for emotional state.
- *Combined features*: As discussed earlier, we explicitly combined some of the above features. For example, we first concatenate all lexicon and hashtag categories found in a given tweet—if the tweet contains both the general hashtag of "obama" and "romney", a binary feature "Hashtag_general_ obama_romney" takes the value of 1.

Results: Table 10 shows the results on this task. Overall, the system obtains an F-measure of 58.30. The table also shows baselines calculated just as described earlier for the emotional state category. We added results for an additional baseline, *rule-based system*, here that chose the stimulus to be: Obama if the tweet had the terms *obama* or *#obama*; Romney if the tweet had the terms *romney* or *#romney*; Republicans if the tweet had the terms *republican, republicans,* or *#republicans*; Democrats if the tweet had the terms *democrats, democrat,* or #democrats; and Campaign if the tweet had the terms *#election* or *#campaign*. If two or more of the above rules were triggered in the same tweet, then a label was chosen at random. This rule-based system obtained an F-score of only 48.62, showing that there were many tweets where key words alone were not sufficient to disambiguate the true stimulus. Observe that the SVM-based automatic system performs markedly better than the majority baseline and also the rule-based system baseline.

4.2. Distinctions between emotion and purpose

The task of detecting purpose is related to sentiment and emotion classification. Intuitively, the three broad categories of purpose, 'oppose', 'favor', and 'other', roughly correspond to negative, positive, and objective sentiment. Also, some fine-grained categories seem to partially correlate with emotions. For example, when angry, a person vents. When overcome with admiration, a person praises the object of admiration.

Since the tweets are annotated for both emotion and purpose, we can investigate the relationship between the two. Fig. 6 shows the percentage of tweets pertaining to different categories of emotion and purpose. The tweets with the purpose 'favor' ('support' and 'praise') mainly convey the emotions of admiration, anticipation, and joy. On the other hand, the tweets with the purpose 'oppose' ('disagree', 'criticize', etc.) are mostly associated with negative emotions such as dislike, anger, and disgust. The purpose 'to praise, admire, or appreciate' is highly correlated with the emotion admiration.

Table 10
Stimulus detection task: overall results on the 8-class task.

	P	R	F
Random baseline	16.45	20.87	18.39
Majority baseline	34.45	38.00	36.14
Automatic rule-based system	43.47	55.15	48.62
Automatic SVM system	57.30	59.32	58.30
Human performance	82.87	81.36	82.11

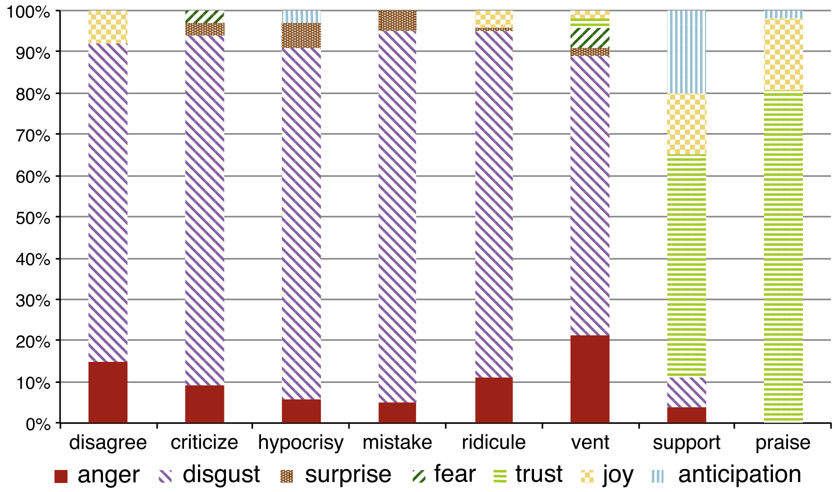

Fig. 6. Percentage of different purpose tweets pertaining to different emotions.

Note that most of the tweets with the purpose ‘to point out hypocrisy’, ‘to point out mistake’, ‘to disagree’, ‘to ridicule’, ‘to criticize’, and even many instances of ‘to vent’ are associated with the emotion dislike. Thus, a system that only determines emotion and not purpose will fail to distinguish between these different categories of purpose. It is possible for people to have the same emotion of dislike and react differently: either by just disagreeing, pointing out the mistake, criticizing, or resorting to ridicule.

4.3. Automatically identifying purpose

To automatically classify tweets into eleven categories of purpose (Table 5), we trained a Support Vector Machine (SVM) classifier. The eleven categories were assumed to be mutually exclusive, i.e., each tweet was classified into exactly one category. In the second set of experiments, the eleven fine-grained categories were combined into 3 coarse-grained – ‘oppose’, ‘favor’, and ‘other’ – as was described earlier. In each experiment, ten-fold stratified cross-validation was repeated ten times, and the results were averaged. Paired t-test was used to confirm the significance of the results. We used the LibSVM package (Chang & Lin, 2011b) with linear kernel and default parameter settings. Parameter C was chosen by cross-validation on the training portion of the data (i.e., the nine training folds).

The gold labels were determined by strong majority voting. Tweets with less than 2 annotations or with no majority labels were discarded. Thus, the dataset consisted of 1072 tweets for the 11-category task, and 1672 tweets for the 3-category task. The tweets were normalized by replacing all URLs with http://someurl and all userids with @someuser. The tweets were tokenized and tagged with parts of speech using the Carnegie Mellon University Twitter NLP tool (Gimpel et al., 2011).

4.3.1. A basic system for purpose classification

Features: Each tweet was represented as a feature vector with the following groups of features.

- ngrams: presence of ngrams (contiguous sequences of 1, 2, 3, and 4 tokens), skipped ngrams (ngrams with one token replaced by *), character ngrams (contiguous sequences of 3, 4, and 5 characters);
- POS: number of occurrences of each part-of-speech;
- word clusters: presence of words from each of the 1000 word clusters provided by the Twitter NLP tool (Gimpel et al., 2011). These clusters were produced with the Brown clustering algorithm on 56 million English-language tweets. They serve as alternative representation of tweet content, reducing the sparcity of the token space;
- all-caps: the number of words with all characters in upper case;
- NRC Emotion Lexicon:
  – number of words associated with each emotion
  – number of nouns, verbs, etc., associated with each emotion
  – number of all-caps words associated with each emotion
  – number of hashtags associated with each emotion

Table 11
Purpose identification task: accuracy of the automatic classification on 11-category and 3-category problems.

	11-class	3-class
Majority class	26.49	58.07
SVM	43.56	73.91

Table 12
Purpose identification task: per category precision (P), recall (R), and F score of the classification on the 11-category problem. Micro-averaged P, R, and F are equal to accuracy since the categories are mutually exclusive.

Category	# inst.	P	R	F
Favor				
To agree	5	0	0	0
To praise	161	57.59	50.43	53.77
To support	284	49.35	69.47	57.71
Oppose				
To point out hypocrisy	75	30.81	21.2	25.12
To point out mistake	37	0	0	0
To disagree	27	0	0	0
To ridicule	165	31.56	43.76	36.67
To criticize	76	22.87	9.87	13.79
To vent	88	36.06	23.07	28.14
Other				
To provide information	143	45.14	50.63	47.73
None of the above	11	0	0	0
Micro-average		43.56	43.56	43.56

- negation: the number of negated contexts. Following (Pang, Lee, & Vaithyanathan, 2002), we defined a negated context as a segment of a tweet that starts with a negation word (e.g., ‘no’, ‘shouldn’t’) and ends with one of the punctuation marks: ‘,’, ‘.’, ‘:’, ‘;’, ‘!’, ‘?’. A negated context affects the ngram and Emotion Lexicon features: each word and associated with it emotion in a negated context become negated (e.g., ‘not perfect’ becomes ‘not perfect_NEG’, ‘EMOTION_trust’ becomes ‘EMOTION_trust_NEG’). The list of negation words was adopted from Christopher Potts’ sentiment tutorial.[11]
- punctuation: the number of contiguous sequences of exclamation marks, question marks, and both exclamation and question marks;
- emoticons: presence/absence of positive and negative emoticons. The polarity of an emoticon was determined with a simple regular expression adopted from Christopher Potts’ tokenizing script.[12]
- hashtags and elongated word: the number of hashtags and the number of words with one character repeated more than 2 times, e.g. ‘soooo’.

Results: Table 11 presents the results of the automatic classification for the 11-category and 3-category problems. For comparison, we also provide the accuracy of a simple baseline classifier that always predicts the majority class. The percentage of error reduction over the baseline is 23.22% for the 11-category classification and 37.78% for the 3-category classification.

Table 12 shows the classification results broken-down by category. As expected, the categories with larger amounts of labeled examples (‘to praise’, ‘to support’, ‘to provide information’) have higher results. However, for one of the higher frequency categories, ‘to ridicule’, the F-score is relatively low. This category incorporates irony, sarcasm, and humor, the concepts that are hard to recognize, especially in a very restricted context of 140 characters. The four low-frequency categories (‘to agree’, ‘to point out mistake or blunder’, ‘to disagree’, ‘none of the above’) did not have enough training data for the classifier to build adequate models. The categories within ‘oppose’ are more difficult to distinguish among than the categories within ‘favor’. However, for the most part this can be explained by the larger number of categories (6 in ‘oppose’ vs. 3 in ‘favor’) and, consequently, smaller sizes of the individual categories.

In the next set of experiments, we investigated the usefulness of each feature group for the task. We repeated the above classification process, each time removing one of the feature groups from the tweet representation. Table 13 shows the results of these ablation experiments for the 11-category and 3-category problems. In both cases, the most influential features were found to be ngrams and emotion lexicon features.

[11] http://sentiment.christopherpotts.net/lingstruc.html.

[12] http://sentiment.christopherpotts.net/tokenizing.html.

Table 13
Purpose identification task: accuracy of classification with one of the feature groups removed. Numbers in bold represent statistically significant difference with the accuracy of the 'all features' classifier (first line) with 95% confidence.

Experiment	11-class	3-class
All features	43.56	73.91
All – ngrams	**39.51**	**71.02**
All – NRC emotion lexicon	**42.27**	**72.21**
All – parts of speech	**42.63**	**73.55**
All – word clusters	**43.24**	**73.24**
All – negation	**43.18**	**73.36**
All – (all-caps, punctuation, emoticons, hashtags)	43.38	73.87

4.3.2. Adding features pertaining to hundreds of fine emotions

Since the emotion lexicon had a significant impact on the results, we further created a wide-coverage tweet-specific lexical resource following on work by Mohammad (2012b). Mohammad (2012b) showed that emotion-word hashtagged tweets are a good source of labeled data for automatic emotion processing. Those experiments were conducted using tweets pertaining to the six Ekman emotions because labeled evaluation data exists for only those emotions. However, a significant advantage of using hashtagged tweets is that we can collect large amounts of labeled data for any emotion that is used as a hashtag by tweeters. Thus we polled the Twitter API and collected a large corpus of tweets pertaining to a few hundred emotions.

We used a list of 585 emotion words compiled by Zeno G. Swijtink as the hashtagged query words.[13] Note that we chose not to dwell on the question of whether each of the words in this set is truly an emotion or not. Our goal was to create and distribute a large set of emotion-labeled data, and users are free to choose a subset of the data that is relevant to their application.

Given a dataset of sentences and associated emotion labels (emotion-word hashtags), we computed the *Strength of Association (SoA)* between a word w and an emotion e to be:

$$SoA(w,e) = PMI(w,e) - PMI(w,\neg e) \tag{1}$$

where PMI is the pointwise mutual information.

$$PMI(w,e) = \log_2 \frac{freq(w,e) * N}{freq(w) * freq(e)} \tag{2}$$

where $freq(w,e)$ is the number of times w occurs in sentences with label e. $freq(w)$ is the frequency of w in the corpus and $freq(e)$ is the total number of tokens in sentences with label e.

N is the number of word tokens in the corpus. Similarly:

$$PMI(w,\neg e) = \log_2 \frac{freq(w,\neg e) * N}{freq(w) * freq(\neg e)} \tag{3}$$

where $freq(w,\neg e)$ is the number of times n occurs in sentences that do not have the label e. $freq(\neg e)$ is the total number of tokens in sentences that do not have the label e. Thus, Eq. 1 is simplified to:

$$SoA(w,e) = \log_2 \frac{freq(w,e) * freq(\neg e)}{freq(e) * freq(w,\neg e)} \tag{4}$$

Since PMI is known to be a poor estimator of association for low-frequency events, we ignore words that occur less than five times in the corpus.

If a word tends to occur more often in a tweet with a particular emotion label, than in a tweet that does not have that label (emotion word hashtag), then that word–emotion pair will have an SoA score that is greater than zero.[14] Consequently, the pairs (word, emotion) that had positive SoA were pulled together into a new word–emotion association resource, that we call *Hashtag Emotion Lexicon*. The lexicon contains around 10,000 words with associations to 585 emotion-word hashtags.

We used the Hashtag Lexicon for classification by creating a separate feature for each emotion-related hashtag, resulting in 585 emotion features. The values of these features were calculated as the sum of the PMI scores between the words in a tweet and the corresponding emotion-related hashtag. Table 14 shows the results of the automatic classification using the

[13] www.sonoma.edu/users/s/swijtink/teaching/philosophy_101/paper1/listemotions.html.

[14] Note that the PMI calculations normalize for number of times the emotion label occurs ($freq(e)$ in Eq. (2)) and the number of times the emotion label does not occur ($freq(\neg e)$ in Eq. 3).

Table 14
Purpose identification task: Accuracy of classification using different lexicons on the 11-class problem. Numbers in bold represent statistically significant difference with the accuracy of the classifier using the NRC Emotion Lexicon (first line) with 95% confidence.

Lexicon	Accuracy
NRC Emotion Lexicon	43.56
Hashtag Lexicon	44.35
Both lexicons	**44.58**

Table 15
A histogram of the number of annotations in response to Questionnaire 1.

Annotations/tweet	# of tweets	# of annotations
1	181	181
2	594	1188
3	1121	3363
4	60	240
⩾5	88	1509
All	2042	6481

new lexical resource. The Hashtag Lexicon significantly improved the performance of the classifier on the 11-category task. Even better results were obtained when both lexicons were employed.[15]

5. Conclusions

Given that social media is playing a growing role in elections world wide, automatically analyzing posts on platforms such as Twitter has a number of applications such as determining support for political parties or candidates, identifying stance of various groups on key electoral issues, determining amount of voter polarization, detecting impact of mass tweets from political parties and Twitter bots in shaping public sentiment, etc. A useful resource in developing such applications is a dataset labeled for various affectual phenomena. Here, for the first time, we collected and annotated a common dataset (2012 US presidential elections tweets) for a number of labels pertaining to sentiment, emotions, style, and purpose. We designed questionnaires specifically for annotation on a crowdsource platform. We analyzed the data to show that electoral tweets are rich in emotions and mostly convey the feelings of the tweeters themselves. The predominant emotion in these tweets is disgust followed by trust. Electoral tweets convey negative emotions twice as often as positive emotions.

We also developed supervised automatic classifiers for detecting emotional state, emotion stimulus, and purpose (or intent) of the tweets. These classifiers used many of the annotations described above for training. The results establish baselines for automatic systems on this new data. We show that even though the purpose classifier benefits from emotion features, emotion detection alone can fail to distinguish between several different types of purpose. For example, the same emotion of disgust can be associated with many different kinds of purpose such as 'to criticize', 'to vent', and 'to ridicule'. Thus, detecting purpose provides information that is not provided simply by detecting sentiment or emotion.

All of the electoral tweets and associated annotations are made freely available.[16] One of our future goals is to use the data to create an automatic system to detect exaggeration. In this paper, we relied only on the target tweet as context. However, it is possible to obtain even better results by modeling user behavior based on multiple past tweets. Another avenue for future research is to compare electoral tweets from different countries, for example, it will be interesting to determine if the distributions of tweets by purpose differ across developed and developing world. We are interested in using purpose-annotated tweets as input in a system that automatically summarizes political tweets. We are also interested in automatically identifying other semantic roles of emotions such as degree, reason, and empathy target (described in Table 1).

Appendix A

The histograms of the annotations from Questionnaires 1 and 2 are shown in Tables 15 and 16. Note that since the tweets with gold questions were used for quality control by CrowdFlower, some tweets were annotated more than three times in Questionnaire 1 and more than 5 times in Questionnaire 2. Only tweets with at least three annotations for Questionnaire 2 and with sufficient agreement among the annotators for emotional state, emotional stimulus, and purpose were used as training and test instances in the automatic classification experiments.

[15] Using the Hashtag Lexicon on the 3-category task did not show any improvement. This is probably because there the information about positive and negative sentiment provides the most gain.

[16] http://www.purl.org/net/PoliticalTweets2012.

Table 16
A histogram of the number of annotations in response to Questionnaire 2.

Annotations/tweet	# of tweets	# of annotations
1	56	56
2	191	382
3	440	1320
4	420	1680
5	108	540
6	13	78
All	1228	4056

References

Alm, E., & Ovesdotter, C. (2008). *Affect in text and speech*. ProQuest.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matoušek & P. Mautner (Eds.), *Text, speech and dialogue. Lecture notes in computer science* (Vol. 4629, pp. 196–205). Berlin/Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-540-74628-7_27.

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. 1204.6441.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics* (Vol. 1, pp. 86–90). Stroudsburg, PA: Association for Computational Linguistics.

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

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

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

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* (pp. 562–570). Stroudsburg, PA, USA: Association for Computational Linguistics.

Caragea, C., McNeese, M., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., et al. (2011). Classifying text messages for the Haiti earthquake. In *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM), Lisbon, Portugal.*

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*.

Chang, C. C., & Lin, C. J. (2011a). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:1–27:27.

Chang, C. C., & Lin, C. J. (2011b). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:1–27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

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, Honolulu, Hawaii* (pp. 793–801).

Choudhury, M. D., Counts, S., & Gamon, M. (2012). Not all moods are created equal! Exploring human emotional states in social media. In *The international AAAI conference on weblogs and social media (ICWSM).*

Chung, J. E., & Mustafaraj, E. (2011). Can collective sentiment expressed on Twitter predict political elections? In W. Burgard & D. Roth (Eds.), *Proceedings of the 25th AAAI conference on artificial intelligence*. California, USA: AAAI Press.

Collier, N., Son, N., & Nguyen, N. (2011). OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics, 2*, S9.

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of Twitter users. In *IEEE third international conference on privacy security risk and trust and IEEE third international conference on social computing* (pp. 192–199). IEEE.

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

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceeding of the 23rd international conference on computational linguistics (COLING)*. Key: citeulike:8637620.

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

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In N. C. C. Chair, K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Genereux, M., & Evans, R. P. (2006). Distinguishing affective states in weblogs. In *AAAI-2006 spring symposium on computational approaches to analysing weblogs, Stanford, California* (pp. 27–29).

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*, 245–288.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., et al. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the annual meeting of the association for computational linguistics.*

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. In *Final projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.*

Golbeck, J., & Hansen, D. (2011). Computing political preference among twitter followers. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1105–1108). New York, NY: ACM.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *ACL (Short Papers)* (pp. 581–586).

Holzman, L. E., & Pottenger, W. M. (2003). Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University. <www.lehigh.edu/~leh7/papers/EmotionClassification.pdf>.

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.

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).* <http://www.cs.indiana.edu/{~}jaskessl/icwsm09.pdf>.

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

Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. M. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the international workshop on semantic evaluation, Dublin, Ireland.*

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

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).

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

Liebrecht, C., Kunneman, F., & van den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 29–37).

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 415–463). US: Springer. http://dx.doi.org/10.1007/978-1-4614-3223-4_13, doi: 10.1007/978-1-4614-3223-4_13.

Ma, C., Prendinger, H., & Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In J. Tao, R. W. Picard (Eds.), *First international conference on affective computing and intelligent interaction (ACII-2005), Beijing, China* (pp. 622–628). <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.81.6625>.

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

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

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.

Mohammad, S. (2012a). 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: Association for Computational Linguistics.

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, 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. M. (2012b). #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* (pp. 246–255). Association for Computational Linguistics, Stroudsburg, PA.

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*.

Naaman, M., Boase, J., & Lai, C. H. (2010). Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM conference on computer supported cooperative work* (pp. 189–192). New York, NY: ACM.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the third international conference on weblogs and social media (ICWSM-09), San Jose, California* (pp. 278–281).

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*.

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

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the conference on empirical methods in natural language processing, Philadelphia, PA* (pp. 79–86).

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

Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. In *Exploring attitude and affect in text: Theories and applications (AAAI spring symposium series)*.

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* (pp. 339–346). Stroudsburg, PA, USA: Association for Computational Linguistics.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*, 130–137.

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* (pp. 38–43). Stroudsburg, PA, USA: Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1859650.1859656>.

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

Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment analysis in twitter. In P. Nakov, T. Zesch (Eds.), *Proceedings of the 8th international workshop on semantic evaluation, Dublin, Ireland*.

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, Jeju, Korea*.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing, Seattle, USA*.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in Twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 841–842). New York, NY: ACM.

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

Strapparava, C., & Valitutti, A. (2004). Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th international conference on language resources and evaluation (LREC-2004), Lisbon, Portugal* (pp. 1083–1086).

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* (pp. 22–30). Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/11940098_3.

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

Tokuhisa, R., Inui, K., & Matsumoto, Y. (2008). Emotion classification using massive examples extracted from the web. *Proceedings of the 22nd international conference on computational linguistics* (Vol. 1, pp. 881–888). Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010b). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Word Journal of the International Linguistic Association*, 178–185. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852>.

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* (pp. 587–592). Washington, DC, USA: IEEE Computer Society.

Wiebe, J., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language: Language resources and evaluation. In *Language resources and evaluation (formerly computers and the humanities)*.

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, Atlanta, Georgia, USA*.

Xu, G., Huang, C. R., & Wang, H. (2013). Extracting chinese product features: Representing a sequence by a set of skip-bigrams. In *Proceedings of the 13th Chinese conference on Chinese lexical semantics* (pp. 72–83). Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-36337-5_9.

Zhang, L., & Liu, B. (2011). Identifying noun product features that imply opinions. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers* (Vol. 2, pp. 575–580). Stroudsburg, PA, USA: Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=2002736.2002849>.

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* (pp. 1462–1470). Stroudsburg, PA, USA: Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1944566.1944733>.

Zhe, X., & Boucouvalas, A. (2002). text-to-emotion engine for real time internet communication text-to-emotion engine for real time internet communication (pp. 164–168). <http://dec.bournemouth.ac.uk/staff/tboucouvalas/staffsJ.pdf>.

Zhu, X., Kiritchenko, S., & Mohammad, S. M. (2014). NRC-Canada-2014: Recent improvements in sentiment analysis of tweets. In *Proceedings of the international workshop on semantic evaluation, Dublin, Ireland*.