Semantic Role Labeling of Emotions in Tweets

Saif M. Mohammad, Xiaodan Zhu, and Joel Martin
National Research Council Canada
Ottawa, Ontario, Canada K1A 0R6
{saif.mohammad,xiaodan.zhu,joel.martin}@nrc-cnrc.gc.ca

Abstract

Past work on emotion processing has focused solely on detecting emotions, and ignored questions such as 'who is feeling the emotion (the experiencer)?' and 'towards whom is the emotion directed (the stimulus)?'. We automatically compile a large dataset of tweets pertaining to the 2012 US presidential elections, and annotate it not only for emotion but also for the experiencer and the stimulus. We then develop a classifier for detecting emotion that obtains an accuracy of 56.84 on an eight-way classification task. Finally, we show how the stimulus identification task can also be framed as a classification task, obtaining an F-score of 58.30.

1 Introduction

Detecting emotions in text has a number of applications including tracking sentiment towards politicians, movies, and products (Pang and Lee, 2008), identifying what emotion a newspaper headline is trying to evoke (Bellegarda, 2010), developing more natural text-to-speech systems (Francisco and Gervás, 2006), detecting how people use emotion-bearing-words and metaphors to persuade and coerce others (for example, in propaganda) (Kővecses, 2003), tracking response to natural disasters (Mandel et al., 2012), and so on. With the rapid proliferation of microblogging, there is growing amount of emotion analysis research on newly available datasets of Twitter posts (Mandel et al., 2012; Purver and Battersby, 2012; Mohammad, 2012b). However, past work has focused solely on detecting emotional state. It has ignored questions such as 'who is feeling the emotion (the experiencer)?' and 'towards whom is the emotion directed (the stimulus)?'.

In this paper, we present a system that analyzes tweets to determine who is feeling what emotion, and towards whom. We use tweets from the 2012 US presidential elections as our dataset, since we expect political tweets to be particularly rich in emotions. Further, the dataset will be useful for applications such as determining political alignment of tweeters (Golbeck and Hansen, 2011; Conover et al., 2011b), identifying contentious issues (Maynard and Funk, 2011), detecting the amount of polarization in the electorate (Conover et al., 2011a), and so on.

Detecting the who, what, and towards whom of emotions is essentially a semantic role-labeling problem (Gildea and Jurafsky, 2002). The semantic frame for 'emotions' in FrameNet (Baker et al., 1998) is shown in Table 1. In this work, we focus on the roles of *Experiencer, State,* and *Stimulus*. Note, however, that the state or emotion is often not explicitly present in text. Other roles such as *Reason, Degree,* and *Event* are also of significance, and remain suitable avenues for future work.

We automatically compile a large dataset of 2012 US presidential elections using a small number of hand-chosen hashtags. Next we annotate the tweets for Experiencer, State, and Stimulus by crowdsourcing to Amazon's Mechanical Turk.[1] We analyze the annotations to determine the distributions of different types of roles, and show that the dataset is rich in emotions. We develop a classifier for emotion detection that obtains an accuracy of 56.84. We find that most of the tweets express emotions of the tweeter, and only a few are indicative of the emotions of someone else. Finally, 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. Our supervised classifier obtains an F-score of 58.30 on this task.

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

Table 1: The FrameNet frame for emotions. The three roles investigated in this paper are shown 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.
Manner	Any way the Experiencer experiences the Stimulus which is not covered by more specific frame elements.
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.

2 Related Work

Our work here is related to emotion analysis, semantic role labeling (SRL), and information extraction (IE).

Much of the past work on emotion detection focuses on emotions argued to be the most basic. For example, Ekman (1992) proposed six basic emotions—joy, sadness, anger, fear, disgust, and surprise. Plutchik (1980) argued in favor of eight—Ekman's six, surprise, and anticipation. Many of the automatic systems use affect lexicons pertaining to these basic emotions such as the NRC Emotion Lexicon (Mohammad and Turney, 2010), WordNet Affect (Strapparava and Valitutti, 2004), and the Affective Norms for English Words.[2] Affect lexicons are lists of words and associated emotions.

Emotion analysis techniques have been applied to many different kinds of text (Mihalcea and Liu, 2006; Genereux and Evans, 2006; Neviarouskaya et al., 2009; Mohammad, 2012a). More recently there has been work on tweets as well (Bollen et al., 2011; Tumasjan et al., 2010; Mohammad, 2012b). Bollen et al. (2011) measured tension, depression, anger, vigor, fatigue, and confusion in tweets. Tumasjan et al. (2010) study Twitter as a forum for political deliberation. Mohammad (2012b) 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.

Semantic role labeling (SRL) identifies semantic arguments and roles with regard to a predicate in a sentence (Gildea and Jurafsky, 2002; Màrquez et al., 2008; Palmer et al., 2010). More recently, there has also been some work on semantic role labeling of tweets for verb and nominal predicates (Liu et al., 2012; Liu et al., 2011). There exists work on extracting opinions and the topics of opinions, however most of it if focused on opinions about product features (Popescu and Etzioni, 2005; Zhang et al., 2010; Kessler and Nicolov, 2009). For example, (Kessler and Nicolov, 2009) identifies semantic relations between sentiment expressions and their targets for car and digital-camera reviews. However, there is no work on semantic role labeling of emotions in tweets. We use many of the ideas developed in the sentiment analysis work and apply them to detect the stimulus of emotions in the electoral tweets data.

Our work here is also related to template filling in information extraction (IE), for example as defined in MUC (Grishman, 1997), which extracts information (entities) from a document to fill out a pre-defined template, such as the date, location, target, and other information about an event.

3 Challenges of Semantic Role Labeling of Emotions in Tweets

Semantic role labeling of emotions in tweets poses certain unique challenges. Firstly, there are many differences between tweets and linguistically well-formed texts, such as written news (Liu et al., 2012; Ritter et al., 2011). Tweets are often less well-formed—they tend to be colloquial, have misspellings, and have non-standard tokens. Thus, methods depending heavily on deep language understanding such as syntactic parsing (Kim and Hovy, 2006) are less reliable.

[2] http://www.purl.org/net/NRCEmotionLexicon
http://csea.phhp.ufl.edu/media/anewmessage.html

Secondly, in a traditional SRL system, an argument frame is a cohesive structure with strong dependencies between the arguments. Thus it is often beneficial to develop joint models to identify the various elements of a frame (Toutanova et al., 2005). However, these assumptions are less viable when dealing with emotions in tweets. For example, there is no reason to believe that people with a certain name will have the same emotions towards the same entities. On the other hand, if we make use of information beyond the target tweet to independently identify the political leanings of a person, then that information can help determine the person's emotions towards certain entities. However, that is beyond the scope of this paper. Thus we develop independent classifiers for identifying experiencer, state, and stimulus.

Often, the goal in SRL and IE template filling is the labeling of text spans in the original text. However, emotions are often not explicitly stated in text. Thus we develop a system that assigns an emotion to a tweet even though that emotion is not explicitly mentioned. The stimulus of the emotion may also not be mentioned. Consider *Happy to see #4moreyears come into reality*. The stimulus of the emotion joy is *to see #4moreyears come into reality*. However, the tweet essentially conveys the tweeter's joy towards Barack Obama being re-elected as president. One may argue that the true stimulus here is Barack Obama. Thus it is useful to normalize mentions and resolve the coreference, for example, all mentions of *Barack H. Obama*, *Barack, Obama,* and *#4moreyears* should be directed to the same entity. Thus, we *ground* (in the same sense as in *language grounding*) the emotional arguments to the predefined entities. Through our experiments we show the target of an emotion in political tweets is often one among a handful of entities. Thus we develop a classifier to identify which of these pre-chosen entities is the stimulus in a given tweet.

4 Data Collection and Annotation

4.1 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.

Table 2: Query terms used to collect tweets pertaining to the 2012 US presidential elections.

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

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 will 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.[3] This step also helps discard very short tweets and tweets with a large proportion of misspelled words. Since we were interested in determining the source and target of emotions in tweets, we decided to focus on original tweets as opposed to retweets. 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.

4.2 Annotating Emotions by Crowdsourcing

We used Amazon's Mechanical Turk service to crowdsource the annotation of the electoral tweets. We randomly selected about 2,000 tweets, each by a different Twitter user. We set up two questionnaires on Mechanical Turk for the tweets. The first questionnaire was used to determine the number of emotions in a tweet and also whether the tweet was truly relevant to the US politics.

[3] www.gutenberg.org/ebooks/10681

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.

Q2. 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.

These questionnaires are called *HITs* (Human Intelligence Tasks) in Mechanical Turk parlance. 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, dozens of annotators contribute assignments for the 2,042 tweets.

The tweets that were marked as having one emotion were chosen for annotation by the Questionnaire 2. We requested responses from at least five annotators for each of these HITs. Below is an example:

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.

- anger or annoyance or hostility or fury
- anticipation or expectancy or interest
- disgust or dislike
- fear or apprehension or panic or terror
- joy or happiness or elation
- sadness or gloominess or grief or sorrow
- surprise
- trust or like

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.

	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

Q3. Towards whom or what?

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 3 as shown below:

Q3. What 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

4.3 Annotation Analyses

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 more than two standard deviations away from the mean. All annotations by these annotators were discarded.

We determine whether a tweet is to be assigned a particular category based on strong majority vote. That is, a tweet belongs to category X if it was annotated by at least three annotators and only if at least half of the annotators agreed with each other. Percentage of tweets in each of the five categories of Q1 are 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: Questionnaire 2: Percentage of tweets in the categories of Q2. 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.

Emotion	Percentage of tweets
anger	7.41
anticipation	5.01
disgust	**47.75**
fear	1.98
joy	6.58
sadness	0.83
surprise	6.37
trust	24.03
all	100.00

Responses to Q2 showed that a large majority (95.56%) of the tweets were relevant to US politics and elections. This shows that 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.

Responses to Q1 of Questionnaire 2 revealed that in the vast majority of the cases (99.825%), the tweets contains emotions of the tweeter. The data did include some tweets that referred to emotions of others such as Romney, GOP, and president, but these instances are rare. Tables 4 and 5 give the distributions of the various options for Questions 2, and 3 of Questionnaire 2. Table 4 shows that disgust (49.32%) is by far the most dominant emotion in the tweets of 2012 US presidential elections. The next most prominent emotion is that of trust (23.73%). About 61% of the tweets convey negative emotions towards someone or something. Table 5 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%.

4.3.1 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 6 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 6 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 experiencer, state, and stimulus 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 5: Questionnaire 2: Percentage of tweets in the categories of Q3. A tweet belongs to category X if it is annotated with X more often than all other categories combined. There were 973 such tweets.

Whom	Percentage of tweets
Barack Obama and/or Joe Biden	**29.90**
Mitt Romney and/or Paul Ryan	24.87
Some other individual	5.03
Democratic party, democrats, or DNC	2.46
Republican party, republicans, or RNC	8.42
Some other institution	1.23
Election campaign or process	4.93
The target is not specified in the tweet	1.95
None of the above	21.17
all	100.00

Table 6: 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	96.76	0.974
Questionnaire 2:		
Q1	52.95	0.731
Q2	59.59	0.736
Q3	44.47	0.641

4.4 Access to the data

All of the data is made freely available through the first author's website:

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

It includes: (1) the complete set of tweets collected from the Twitter API with hashtags shown in Table 2, (2) the subset of English tweets, (3) Questionnaires 1 and 2, (4) and tweets annotated as per Questionnaires 1 and 2.

5 Automatically Detecting Semantic Roles of Emotions in Tweets

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

Due to the unique challenges of semantic role labeling of emotions in tweets described earlier in the paper, 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 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 stratified ten-fold cross validation. We found that a simple linear kernel yielded the best performance. We used the LibSVM package (Chang and Lin, 2011).

As mentioned earlier, there is fair amount of work on emotion detection in non-tweet texts (Boucouvalas, 2002; Holzman and Pottenger, 2003; Ma et al., 2005; John et al., 2006; Mihalcea and Liu, 2006; Genereux and Evans, 2006; Aman and Szpakowicz, 2007; Tokuhisa et al., 2008; Neviarouskaya et al., 2009) as well as on tweets (Kim et al., 2009; Tumasjan et al., 2010; Bollen et al., 2011; Mohammad, 2012b; Choudhury et al., 2012; Wang et al., 2012). In the experiments below we draw from various successfully used features described in these papers. More specifically, the system we use builds on the classifier and features used in two previous systems: (1) the system described in (Mohammad, 2012b) which was shown to perform significantly better than some other previous systems on the news paper headlines corpus and the system described in (Mohammad et al., 2013) which ranked first (among 44 participating teams) in a 2013 SemEval competition on detecting sentiment in tweets).

The goal of the experiments in this section is to apply a state-of-the art emotion detection system on the electoral tweets data. We want to set up baseline performance for emotion detection on this new dataset and also validate the data by showing that automatic classifiers can obtain results that are greater than random and majority baselines. In Section 5.2, we apply the SVM classifier and various features for the first time on the task of detecting the stimulus of an emotion in tweets. In each experiment, we report results of ten-fold stratified cross-validation.

5.1 Detecting emotional state

5.1.1 Features

We included the following features for detecting emotional state in tweets.

Word n-grams: We included unigrams (single words) and bigrams (two-word sequences) into our feature set. 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 final character repeated 3 or more times (*soooo, mannnnnn*, etc). (Elongated words have been used similarly in (Brody and Diakopoulos, 2011).)

Emoticons: presence/absence of positive and negative emoticons. The emoticon and its polarity were determined through a regular expression adopted from Christopher Potts' tokenizing script.[4]

Emotion Lexicons: We used the NRC word–emotion association lexicon (Mohammad and 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. We also counted the number of words with regard to the Osgood's (Osgood et al., 1957) semantic differential categories ($LexOsg$) built for Wordnet ($LexOsg\_wn$) and General Inquirer ($LexOsg\_gi$). To reduce noise, we only considered the words that have an adjective or adverb sense in Wordnet.

Negation features: We examined tweets to determine whether they contained negators such as *no, not,* and *shouldn't*. An additional feature determined whether the negator was located close to an

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

Table 7: Results for emotion detection.

	Accuracy
random baseline	30.26
majority baseline	47.75
automatic SVM system	56.84
upper bound	69.80

Table 8: The accuracies obtained with one of the feature groups removed. The number in brackets is the difference with the *all features* score. The biggest drop is shown in bold.

Experiment	**Accuracy**	Difference from all features
all features	56.84	0
all - ngrams	53.35	**-3.49**
all - word ngrams	54.44	-2.40
all - char. ngrams	56.32	-0.52
all - lexicons	54.34	-2.50
all - manual lex.	55.17	-1.67
all - auto lex.	55.38	-1.46
all - negation	55.80	-1.04
all - encodings (elongated words, emoticons, punctns., uppercase)	56.82	-0.02

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.[5]

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. Also, if a tweet contained more than one joy word and no other emotion words, then the feature $LexEmo\_joy\_only$ was triggered.

5.1.2 Results

Table 7 shows the results. We included two baselines here: 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 relative high.

The automatic system obtained by the classifier in identifying the emotions (56.84), which is significantly higher than the majority baseline. It should be noted that the highest scores in the SemEval 2013 task of detecting sentiment analysis of tweets was around 69% (Mohammad 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.

The upper bound of the task here is not 100%—human annotators do not always agree with each other. To estimate the upper bound we can expect an automatic system to achieve, for each tweet we randomly sampled an human annotation from its multiple annotations and treated it as a system output. 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 results table shows this upper bound.

Table 8 shows 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 from ngrams come through word ngrams, but character ngrams provide small gains as well. Both the manual and automatic sentiment lexicons were found to be useful to a similar degree. Paying attention to negation was also beneficial, whereas 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.

5.2 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 categories defined in Table 5. We believe that a similar approach is also possible in other

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

Table 9: Results for detecting stimulus.

	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
upper bound	82.87	81.36	82.11

domains such as natural disaster tweets and epidemic surveillance tweets. We perform a ten-fold stratified cross-validation.

5.2.1 Features

We used the features 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 Table 5. For example, the Republican party may be called as "gop" or "Grand Old Party"; all such words or phrases are all 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.

5.2.2 Results

Table 9 shows the results obtained by the system. Overall, the system obtains an F-measure of 58.30. The table also shows upper-bound and 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 are triggered in the same tweet, then a label is chosen at random. This rule-based system based on hand-chosen features obtains an F-score of 48.62, showing that there are sufficiently many tweets where key words alone are 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.

6 Conclusions and Future Work

In this paper, we framed emotion detection as a semantic role labeling problem, focusing not just on emotional state but also on experiencer and stimulus. We chose tweets about the 2012 US presidential elections as our target domain. We automatically compiled a large dataset of these tweets using hashtags, and annotated them first for presence of emotions, and then for the different semantic roles of emotions. All of the data is made freely available.

We found that a large majority of these tweets (88.1%) carry some emotional attitude towards someone or something. Further, tweets that convey disgust are twice as prevalent than those that convey trust. We found that most tweets express emotions of the tweeter themselves, and the stimulus is often one among a few handful of entities. We developed a classifier for emotion detection that obtained an accuracy of 56.84 on an eight-way classification task. Finally, we showed how the stimulus identification task can be framed as a classification task in which our system outperforms competitive baselines.

Our future work involves exploring the use of more tweets from the same user to determine their political leanings, and use that as an additional feature in emotion detection. We are also interested in automatically identifying other semantic roles of emotions such as degree, reason, and empathy target (described in Table 1). We believe that a more sophisticated sentiment analysis applications and a better understanding of affect require the determination of semantic roles of emotion.

References

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In Vclav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer Berlin / Heidelberg.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA. Association for Computational Linguistics.

Jerome Bellegarda. 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.

Johan Bollen, Alberto Pepe, and Huina Mao. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain.

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

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 562–570, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 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)*.

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

Michael D Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011b. Predicting the political alignment of Twitter users. In *IEEE Third International Conference on Privacy Security Risk and Trust and IEEE Third International Conference on Social Computing*, pages 192–199. IEEE.

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

Virginia Francisco and Pablo Gervás. 2006. Automated mark up of affective information in english texts. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, pages 375–382. Springer Berlin / Heidelberg.

Michel Genereux and Roger Evans. 2006. Distinguishing affective states in weblogs. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 27–29, Stanford, USA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Jennifer Golbeck and Derek Hansen. 2011. Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1105–1108, New York, NY. ACM.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *SCIE*, pages 10–27.

Lars E. Holzman and William M. Pottenger. 2003. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University.

David John, Anthony C. Boucouvalas, and Zhe Xu. 2006. Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pages 183–188, Anaheim, CA. ACTA Press.

Jason S. Kessler and Nicolas Nicolov. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2009)*.

S. Kim and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8.

Elsa Kim, Sam Gilbert, Michael J. Edwards, and Erhardt Graeff. 2009. Detecting sadness in 140 characters: Sentiment analysis of mourning Michael Jackson on twitter.

Zoltán Kővecses. 2003. *Metaphor and Emotion: Language, Culture, and Body in Human Feeling (Studies in Emotion and Social Interaction)*. Cambridge University Press.

X. Liu, K. Li, M. Zhou, and Z. Xiong. 2011. Enhancing semantic role labeling for tweets using self-training. In *AAAI*.

X. Liu, Z. Fu, F. Wei, and M. Zhou. 2012. Collective nominal semantic role labeling for tweets. In *AAAI*.

Chunling Ma, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion estimation and reasoning based on affective textual interaction. In J. Tao and R. W. Picard, editors, *First International Conference on Affective Computing and Intelligent Interaction (ACII-2005)*, pages 622–628, Beijing, China.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. 2012. A demographic analysis of online sentiment during Hurricane Irene. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 27–36, Stroudsburg, PA. Association for Computational Linguistics.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Diana Maynard and Adam Funk. 2011. Automatic detection of political opinions in tweets. *gateacuk*, 7117:81–92.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 139–144. AAAI Press.

Saif M. Mohammad and Peter D. Turney. 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.

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

Saif Mohammad. 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*, pages 587–591, Montréal, Canada.

Saif M. Mohammad. 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*, SemEval '12, pages 246–255, Stroudsburg, PA.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Third International Conference on Weblogs and Social Media*, pages 278–281, San Jose, California.

C.E. Osgood, Suci G., and P. Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

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

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346, Stroudsburg, PA, USA.

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

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 482–491, Stroudsburg, PA. Association for Computational Linguistics.

A. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 881–888, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2005. Joint learning improves semantic role labeling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 589–596, Stroudsburg, PA. Association for Computational Linguistics.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with Twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, pages 178–185.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing*, SOCIALCOM-PASSAT '12, pages 587–592, Washington, DC, USA. IEEE Computer Society.

Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. 2010. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1462–1470, Stroudsburg, PA, USA. Association for Computational Linguistics.