Identifying Purpose Behind Electoral Tweets

Saif M. Mohammad, Svetlana Kiritchenko, and Joel Martin
National Research Council Canada
Ottawa, Ontario, Canada K1A 0R6
{saif.mohammad,svetlana.kiritchenko,joel.martin}@nrc-cnrc.gc.ca

ABSTRACT

Tweets pertaining to a single event, such as a national election, can number in the hundreds of millions. Automatically analyzing them is beneficial in many downstream natural language applications such as question answering and summarization. In this paper, we propose a new task: identifying purpose behind electoral tweets—why do people post election-oriented tweets? We show that identifying purpose is related to sentiment and emotion detection, but yet significantly different. Detecting purpose has a number of applications including detecting the mood of the electorate, estimating the popularity of policies, identifying key issues of contention, and predicting the course of events. We create a large dataset of electoral tweets and annotate a few thousand tweets for purpose. We develop a system that automatically classifies electoral tweets as per their purpose, obtaining an accuracy of 44.58% on an 11-class task and an accuracy of 73.91% on a 3-class task (both accuracies well above the most-frequent-class baseline). We also show that resources developed for emotion detection are helpful for detecting purpose.

1. INTRODUCTION

The number of tweets pertaining to a single event or topic such as a national election, a natural disaster, or gun control laws, 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 tweets is beneficial in many downstream natural language applications such as question answering and summarization.

An important facet in understanding tweets is the question of 'Why?', that is, what is the purpose or intent of the tweet? There has been some prior work in this regard [1, 24, 33], however, they have focused on the general motivations and reasons for tweeting. For example, Naaman et al. [24] proposed the categories of: information sharing, self promotion, opinions, statements, me now, questions, presence maintenance, anecdote (me), and anecdote (others). On the other hand, the dominant reasons for tweeting vary when tweeting about specific topics and events. For example, the reasons for tweeting in national elections are very different from the reasons for tweeting during a natural disaster, such as an earthquake.

There is growing interest in analyzing political tweets in particular because of a number of applications such as determining political alignment of tweeters [13, 9], identifying contentious issues and political opinions [19], detecting the amount of polarization in the electorate [10], and so on. There is even a body of work claiming that analyzing political tweets can help predict the outcome of elections [4, 37]. However, that claim is questioned by more recent work [2].

In this paper, we propose the task of identifying the purpose behind electoral tweets. For example, some tweets are meant to criticize, some to praise, some to express disagreement, and so on. Determining the purpose behind electoral tweets can help many applications such as those listed above. There are many reasons why people criticize, praise, etc, but that is beyond the scope of this paper. For discussions on user satisfaction from tweets we refer the reader to work by Liu, Cheung, and Lee [18].

First, we automatically compile a dataset of electoral tweets using a few hand-chosen hashtags. We choose the 2012 US presidential elections as our target domain. We develop a questionnaire to annotate tweets for purpose by crowdsourcing. We analyze the annotations to determine the distributions of different kinds of purpose. We show that emotion detection alone can fail to distinguish between several different types of purpose. For example, the same emotion of dislike 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 obtainable simply by detecting sentiment or emotion.

Next, we develop 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 emoticons, 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

WISDOM'13, August 11 2013, Chicago, USA.

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. We publicly release all the data created as part of this project: about 1 million original tweets on the 2012 US elections, about 2,000 tweets annotated for purpose, about 1,200 tweets annotated for emotion, and the new emotion lexicon.[1]

This paper is organized as follows. We begin with related work (Section 2). We then describe how we collected and annotated the data (Sections 3.1 and 3.2). Section 3.3 gives an analysis of the annotations including distributions of various kinds of purpose, inter-annotator agreement, and confusion matrices. In Section 3.4, we flesh out the partial correlation and the distinction between purpose and affect. In Section 4, we first present a basic system to classify tweets by purpose (Section 4.1), and then we describe how we created an emotion resource pertaining to hundreds of emotions and used it to further improve performance of the basic system (Section 4.2) We present concluding remarks in Section 5.

2. RELATED WORK

There exists considerable work on tweet classification by topic [32, 17, 25]. Some of the classification work that comes close to identifying purpose is described below. Alhadi et al. [1] annotated 1000 tweets into the categories of social interaction with people, promotion or marketing, share resources, give or require feedback, broadcast alert/urgent information, require/raise funding, recruit worker, and express emotions. Naaman et al. [24] organized 3379 tweets into the categories of information sharing, self promotion, opinions, statements, me now, questions, presence maintenance, anecdote (me), and anecdote (others). Sankaranarayanan et al. [33] built a system to identify tweets pertaining to breaking news. Sriram et al. [34] annotated 5407 tweets into news, events, opinions, deals and private messages.

Tweet categorization work within a particular domain includes that by Collier, Son, and Nguyen [8], where flu-related tweets were classified into avoidance behavior, increased sanitation, seeking pharmaceutical intervention, wearing a mask, and self reported diagnosis, and work by Caragea et al. [5], where earthquake-related tweets were classified into medical emergency, people trapped, food shortage, water shortage, water sanitation, shelter needed, collapsed structure, food distribution, hospital/clinic services, and person news.

To the best of our knowledge, there is no work yet on classifying electoral or political tweets into sub-categories. As mentioned earlier, there exists work on determining political alignment of tweeters [13, 9], identifying contentious issues and political opinions [19], detecting the amount of polarization in the electorate [10], and detecting sentiment in political tweets [4, 7].

Sentiment classification of general (non-domain) tweets has received much attention [26, 14, 16]. Beyond simply positive and negative sentiment, some recent work also classifies tweets into emotions [15, 20, 31, 36]. Much of this work focused on emotions argued to be the most basic. For example, Ekman [11] proposed six basic emotions—joy, sadness, anger, fear, disgust, and surprise. Plutchik [30] argued in favor of eight—Ekman's six, trust, and anticipation. There is less work on complex emotions, such as work by Pearl and Steyvers [29] that focused on politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief.

Many of the automatic emotion classification systems use affect lexicons such as the NRC emotion lexicon [22, 23], WordNet Affect [35], and the Affective Norms for English Words.[2] Affect lexicons are lists of words and associated emotions and sentiments. We will show that affect lexicons are helpful for detecting purpose behind tweets as well.

Table 1: 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

3. DATA COLLECTION AND ANNOTATION OF PURPOSE

In the subsections below we describe how we collected tweets posted during the run up to the 2012 US presidential elections and how we annotated them for purpose by crowdsourcing.

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 1 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 were the names of the two presidential candidates. Further, the probability that these words were used to refer to someone other than the presidential candidates 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 will make freely available to the research community.[3] 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.

[1]Email Saif Mohammad: saif.mohammad@nrc-cnrc.gc.ca.

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

[3]Note that Twitter imposes restrictions on direct distribution of tweets, but allows the distribution of tweet ids. One may download tweets using tweet ids and third party tools, provided those tweets have not been deleted by the people who posted them.

We used the metadata tag "iso_language_code" to identify English tweets. Since this tag does not always correctly reflect the language of the tweet, 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.

Since we were interested in determining the purpose behind the tweets, we decided to focus on original tweets as opposed to retweets. Retweets can easily be identified through the presence of RT, rt, or Rt in the tweet (usually in the beginning of the post). All such tweets were discarded.

3.2 Annotating Purpose by Crowdsourcing

We used Amazon's Mechanical Turk service to crowdsource the annotation of the electoral tweets.[4] We randomly selected about 2,000 tweets, each by a different Twitter user. We asked a series of questions for each tweet. Below is the questionnaire for an example tweet:

Purpose behind US election tweets

Tweet: Mitt Romney is arrogant as hell.

Q1. 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 provide information without emotion
- none of the above

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, about 400 annotators contribute assignments for the 2,042 tweets. The number of assignments completed by the annotators followed a zipfian distribution.

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.

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

Table 2: The histogram of the number of annotations of tweets. 'annotns' is short for annotations.

annotns/tweet	# of tweets	# of annotns
1	181	181
2	594	1188
3	1121	3363
4	60	240
≥5	88	1509
all	2042	6481

We created an initial set of categories of purpose by consultations with colleagues and analysis of a small set of tweets. We further refined the set of categories after a pilot annotation project by removing categories that were not represented in the data and also categories that were confused with others. For example, we removed the category 'to entertain' as it was found to intersect with several other categories.

Observe that we implicitly grouped the final options for Q1 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), *favour* (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.

3.3 Annotation Analyses

The Mechanical Turk annotations were done over a period of one week. For each annotator, and for each question, we calculated the probability with which the annotator agrees 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. Table 2 gives a histogram of the number of annotations of the remaining tweets. There were 1121 tweets with exactly three annotations.

We determined whether a tweet is to be assigned a particular category based on strong majority. That is, a tweet belongs to category X if it is annotated with X more often than all other categories combined. Percentage of tweets in each of the 11 categories of Q1 are shown in Table 3. Observe that the majority category for purpose is 'to support'—26.49% of the tweets were identified as having the purpose 'to support'. Table 4 gives the distributions of the three coarse categories of purpose. Observe, that the political tweets express disagreement (58.07%) much more than support (31.76%).

Table 5 gives the distributions for question 2. Observe that a large majority (95.56%) of the tweets are relevant to US politics and elections. This shows that the hashtags shown earlier in Table 1 are effective in identifying political tweets.

Table 3: Percentage of tweets in each of the eleven categories 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 1072 such tweets in total.

Purpose of tweet	Percentage of tweets
favour	
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

Table 4: Percentage of tweets in each of the three coarse categories 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 1672 such tweets in total. The annotator agreement on the three categories is larger than on eleven categories.

Category	Percentage of tweets
oppose	**58.07**
favour	31.76
other	10.17
all	100.0

3.3.1 Inter-Annotator Agreement

We calculated agreement on the full set of annotations, and not just on the annotations with a strong majority as described in the previous section. One way to gauge the amount of agreement among annotators is to examine the number of times all three annotators agree (majority class size = 3), the number of times two out of three annotators agree (majority class size = 2), and the number of times all three annotators choose different options (majority class size = 1).

Table 6 gives the distributions of the majority classes. Higher numbers for the larger class sizes indicate higher agreement. For example, for 22.4% of the tweets all three annotators gave the same answer for question 1 (Q1). The agreement is much higher if one only considers the coarse categories of 'oppose', 'favour', and 'other'—these numbers are shown in the row marked Q1'. The agreement for question 2 was substantially high. This was expected as it is a relatively straightforward question. The numbers in the table are calculated from tweets with exactly three annotations.

Table 5: Percentage of tweets in each of the two categories of Q2.

Relevance	Percentage of tweets
pertaining to US politics and elections	**95.56**
not pertaining to US politics and elections	4.44
all	100.0

Table 6: Percentage of tweets having majority class size (MCS) of 1, 2, and 3. Note: Q is short for question.

	MCS-1	MCS-2	MCS-3
Q1	29.5	48.1	22.4
Q1'	2.2	31.7	66.1
Q2	0.0	5.7	94.3

Table 7 shows *inter-annotator agreement (IAA)*, for the two questions—the average percentage of times two annotators agree with each other. IAA gives us an understanding of the degree of agreement through a single number. Observe that the agreement is only moderate for the eleven fine categories of purpose (43.58%), but much higher when considering the coarser categories (83.81%).

Another way to gauge agreement is by calculating the average probability with which an annotator picks the majority class. Consider the example below: Each tweet is annotated by 3 different annotators. X annotates 10 tweets. Six of the times, X's answer for Q1 is the answer that has a majority (in case of 3 annotators, this means that at least one other annotator also gave the same answer as X for 6 of the 10 tweets). Thus the probability with which X picks the majority class is 6/10. The last column in Table 7 shows the *average probability of picking the majority class (APMS)* by the annotators (higher numbers indicate higher agreement). Overall, we observe that there is strong agreement between annotators at identifying whether the purpose of a tweet is to oppose, to favour, or something else.

3.3.2 Confusion Matrix

Human annotators may disagree with each other because two or more options may seem appropriate for a given tweet. There also exist tweets where the purpose is unclear. Table 8 shows the confusion matrix for question 1. The rows and columns of the matrix correspond to the eleven options. The value in a particular cell, say for row x and column y, is the number of annotations that were assigned label y even though the majority votes for each of those tweets were for x. The highest number in each row is shown in bold. The cells in the diagonal correspond to the number of instances for which the annotations matched the majority vote. For high agreement, one would want higher numbers in the diagonal, which is what we observe in Table 8.

We can identify options that tend to be confused for each other by noting non-diagonal cells with high values. For example, consider cell r7–c8. The relatively large number indicates that 'to ridicule' is sometimes confused with 'to

Table 8: Confusion Matrix: Question 1 (fine-grained). The value in a particular cell, say for row x and column y, is the number of annotations that were assigned label y even though the majority votes for each of those tweets were for x. The highest number in each row is shown in bold.

		c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
favour												
to agree:	r1	**20**	5	9	2	1	2	0	3	0	4	0
to praise, admire, or appreciate:	r2	0	**291**	61	1	1	5	1	5	4	3	0
to support:	r3	1	43	**565**	5	4	23	7	18	5	22	3
oppose												
to point out hypocrisy or inconsistency:	r4	2	2	14	**123**	15	26	10	64	11	5	0
to point out mistake or blunder:	r5	0	6	16	6	**84**	29	15	46	1	3	0
to disagree:	r6	0	0	5	10	2	**145**	10	5	5	1	0
to ridicule:	r7	3	11	28	9	16	37	**274**	60	15	4	0
to criticize, but none of the above:	r8	1	0	22	8	5	49	30	**227**	9	3	0
to vent:	r9	7	12	35	5	11	37	22	45	**155**	7	1
other												
to provide information without any emotional content:	r10	2	11	39	1	4	8	11	19	8	**259**	4
none of the above:	r11	3	6	10	1	4	5	7	3	6	10	**19**

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

	IAA	APMS
Q1	43.58	0.520
Q1'	83.81	0.855
Q2	96.76	0.974

criticize, but none of the above'. Similarly, we find that 'to point out hypocrisy or inconsistency' and 'to point out mistake or blunder' can also be confused with 'to criticize, but none of the above' (r4–c8 and r5–c8). Note however, that the labels are not confused as strongly in the other direction. For example, tweets that have a purpose of 'to criticize' are not confused as much with 'to point out hypocrisy' (r8–c4). This suggests that the category 'to criticize, but none of the above' serves as a hold-back for other finer-grained categories of 'oppose' and, therefore, is often chosen by annotators for less clear messages. A similar situation occurs in the 'favour' group, where the confusion occurs mostly between a more general category 'to support' and more specific categories 'to agree' and 'to praise, admire, or appreciate'.

Note that in a particular application, one may choose only a subset of the eleven categories that are most relevant. For example, one may combine 'to point out hypocrisy', 'to point out mistake', and 'to criticize, but none of the above' into a single category, and distinguish it from other oppose categories such as 'to disagree' and 'to ridicule'.

Table 9 shows the confusion matrix within the coarse categories of question 1. The confusion between the coarse categories is lower than among the finer categories, but yet there exist instances when 'favour' is confused with 'oppose', and vice versa. Table 10 shows the confusion matrix for question 2. Only a very small number of instances are confused with the wrong option for this question.

Table 9: Confusion Matrix: Question 1' (coarse grained).

		c1	c2	c3
favour:	r1	**941**	136	37
oppose:	r2	75	**1705**	29
other:	r3	40	88	**312**

Table 10: Confusion Matrix: Question 2.

		c1	c2
not pertaining to US politics and elections:	r1	**106**	38
pertaining to US politics and elections:	r2	26	**3193**

3.4 Distinctions between purpose and affect

The task of detecting purpose is related to sentiment and emotion classification. Intuitively, the three broad categories of purpose, 'oppose', 'favour', 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.

To further investigate the relation between purpose and emotion, we annotated a portion of the tweets by crowdsourcing with one of 19 emotions: acceptance, admiration, amazement, anger, anticipation, calmness, disappointment, disgust, dislike, fear, hate, indifference, joy, like, sadness, surprise, trust, uncertainty, and vigilance. Similar to the annotation of purpose, each tweet was annotated by at least two judges, and tweets with no strong majority were discarded.

Table 11 shows the percentage of tweets pertaining to different emotions. Only high-frequency categories of purpose and emotion are shown. As expected, the tweets with the purpose 'favour' mainly convey the emotions of admiration,

Table 11: Percentage of different purpose tweets pertaining to different emotions. Low-frequency categories of purpose and emotion are omitted. The highest number for each category of purpose is shown in bold.

	admiration	anticipation	joy	dislike	disappointment	disgust	anger
favour							
to praise, admire, or appreciate	**67**	4	25				
to support	**33**	21	21	4	2		7
oppose							
to point out hypocrisy or inconsistency				**61**		17	11
to point out mistake or blunder				**77**		15	8
to disagree			14	**43**		14	29
to ridicule			7	**66**		7	18
to criticize, but none of the above				**47**	11	16	16
to vent			4	24	12	8	**36**

anticipation, and joy. On the other hand, the tweets with the purpose 'oppose' 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.

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. DETECTING PURPOSE

In this section, we investigate the usefulness of emotion resources in automatically detecting purpose. We train an automatic classifier over an extensive set of features drawn from those used for sentiment analysis of social media texts [27, 3, 21] as well as emotion features and determine the impact of each feature group on classifier performance.

We used a Support Vector Machine (SVM) classifier as they have been shown to be effective on text categorization tasks and robust on large feature spaces. We used the LibSVM package [6] with linear kernel. Parameter C was chosen by cross-validation on the training portion of the data (i.e., the nine training folds). We first classified the tweets into one of eleven categories of purpose. In a second set of experiments, the eleven fine-grained categories were combined into 3 coarse-grained - 'oppose', 'favour', 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.

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 URLs with http://someurl and userids with @someuser. The tweets were tokenized and tagged with parts of speech using the Carnegie Mellon University Twitter NLP tool [12].

4.1 A Basic System for Purpose Classification

We employed commonly used text classification features such as ngrams, part-of-speech, and punctuations, as well as common Twitter-specific features such as emoticons and hashtags. Additionally, we hypothesized that the purpose of tweets is guided by the emotions of the tweeter. Thus we explored certain emotion features as well. Each tweet was represented with the following groups of features:

- n-grams: presence of n-grams (contiguous sequences of 1, 2, 3, and 4 tokens), skipped n-grams (n-grams with one token replaced by *), character n-grams (contiguous sequences of 3, 4, and 5 characters);
- POS: number of occurrences for each part-of-speech tag;
- word clusters: presence of words from each of the 1000 word clusters provided by the Twitter NLP tool [12]. 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: number of words with all characters in upper case;
- NRC Emotion Lexicon: We used the NRC Emotion Lexicon [22] to incorporate affect features. The lexicon consists of 14,182 words manually annotated with 8 basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and 2 polarities (positive, negative). Each word can have zero, one, or more associated emotions and zero or one polarity. For each tweet we counted:
  - 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
- negation: the number of negated contexts. Following [28], 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 n-gram and Emotion Lexicon features: each word and associated with it emotion in a negated context become

Table 12: Accuracy of the automatic classification on 11-category and 3-category problems. The lower bound is the percentage of the majority class.

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

Table 13: Per category precision (P), recall (R), and F1 score of the classification on the 11-category problem. Micro-averaged P, R, and F1 are equal to accuracy since the categories are mutually exclusive.

category	# inst.	P	R	F1
favour				
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-ave		43.56	43.56	43.56

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.[5]

- 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.[6]
- hashtags: the number of hashtags;
- elongated words: the number of words with one character repeated more than 2 times, e.g. 'soooo'.

Table 12 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.

Table 13 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 F1-score is relatively low. This category incorporates irony, sarcasm, and humour, 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 'favour'. However, for the most part this can be explained by the larger number of categories (6 in 'oppose' vs. 3 in 'favour') and, consequently, smaller sizes of the individual categories.

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

Table 14: 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 - n-grams	**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

Table 15: 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**

We investigated the usefulness of each feature group by repeating the above classification process and each time removing one of the feature groups. Table 14 shows the results of these ablation experiments for the 11-category and 3-category problems. In both cases, the emotion lexicon features were found to be helpful and provided significant gains, second only to the ngram features.

4.2 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 twitter-specific lexical resource following on work by Mohammad [20]. [20] 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.[7] 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 affect-labeled data, and users are free to choose a subset of the data that is relevant to their application. We calculated the pointwise mutual information (PMI) between an emotional hashtag and a word appearing in tweets. The PMI represents a degree of correlation between the word and emotion, with larger scores representing stronger correlations. Consequently, the pairs (word, hashtag) that had positive PMI 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 features in the same way as we did for the NRC Emotion Lexicon. Since the Hashtag Lexicon additionally provides real-valued scores of association, for each tweet, we calculated the sum of these scores instead of simply counting the number of emotion-associated words. Table 15 shows the results. The Hashtag Lexicon improved the performance of the classifier on the 11-category task. Even better results were obtained when both lexicons were employed (the improvement over the NRC Emotion Lexicon is statistically significant)[8].

5. CONCLUSIONS

Tweets are playing a growing role in the public discourse on politics. In this paper, we explored the purpose behind such tweets. Detecting purpose has a number of applications including detecting the mood of the electorate, estimating the popularity of policies, identifying key issues of contention, and predicting the course of events. We compiled a dataset of 1 million tweets pertaining to the 2012 US presidential elections using relevant hashtags. We designed an online questionnaire and annotated a few thousand tweets for purpose via crowdsourcing. We analyzed these tweets and showed that a large majority convey emotional attitude towards someone or something. Further, the number of messages posted to oppose someone or something were almost twice the number of messages posted to offer support.

We developed a classifier to automatically classify electoral tweets as per their purpose. It obtained an accuracy of 44.58% on a 11-class task and an accuracy of 73.91% on a 3-class task (both accuracies well above the most-frequent-class baseline). We found that word–emotion association resources such as the NRC Emotion Lexicon and the Hashtag Emotion Lexicon are helpful for detecting purpose. However, we also showed that emotion detection alone can fail to distinguish between several kinds of purpose. We make all the data created as part of this research freely available.

In this paper, we relied only on the target tweet as context. However, it might be possible to further improve results by modeling user behaviour based on multiple past tweets. We are also interested in using purpose-annotated tweets as input in a system that automatically summarizes political tweets. Finally, we hope that a better understanding of purpose of tweets will help drive the political discourse towards issues and concerns most relevant to the people.

[7] http://www.sonoma.edu/users/s/swijtink/teaching/philosophy_101/paper1/listemotions.htm

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

6. REFERENCES

[1] A. C. Alhadi, S. Staab, and T. Gottron. Exploring User Purpose Writing Single Tweets. In *WebSci'11: Proceedings of the 3rd International Conference on Web Science*, 2011.

[2] D. G. Avello. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" – A Balanced Survey on Election Prediction using Twitter Data. *arXiv*, 1204.6441, 2012.

[3] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of Coling: Poster Volume*, pages 36–44, Beijing, China, August 2010.

[4] A. Bermingham and A. Smeaton. On Using Twitter to Monitor Political Sentiment and Predict Election Results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, Chiang Mai, Thailand, 2011. Asian Federation of Natural Language Processing.

[5] C. Caragea, M. McNeese, A. Jaiswal, G. Traylor, H. Kim, P. Mitra, D. Wu, A. Tapia, C. Giles, J. Jansen, and J. Yen. 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, 2011.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[7] J. E. Chung and E. Mustafaraj. Can Collective Sentiment Expressed on Twitter Predict Political Elections? In W. Burgard and D. Roth, editors, *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, California, USA, 2011. AAAI Press.

[8] N. Collier, N. Son, and N. Nguyen. OMG U got flu? Analysis of Shared Health Messages for Bio-surveillance. *Journal of Biomedical Semantics*, 2(Suppl 5):S9, 2011.

[9] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 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, 2011.

[10] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonc, A. Flammini, and F. Menczer. Political Polarization

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

[11] P. Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, 6(3):169–200, 1992.

[12] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011.

[13] J. Golbeck and D. Hansen. 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, 2011. ACM.

[14] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-Dependent Twitter Sentiment Classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, pages 151–160, 2011.

[15] S. Kim, J. Bak, and A. H. Oh. Do You Feel What I Feel? Social Aspects of Emotions in Twitter Conversations. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2012.

[16] E. Kouloumpis, T. Wilson, and J. Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[17] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter Trending Topic Classification. In *Proceedings of the IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, pages 251–258. IEEE, 2011.

[18] I. L. B. Liu, C. M. K. Cheung, and M. K. O. Lee. *Understanding Twitter Usage: What Drive People Continue to Tweet*, pages 928–939. 2010.

[19] D. Maynard and A. Funk. Automatic Detection of Political Opinions in Tweets. In *The Semantic Web: ESWC 2011 Workshops*, pages 88–99. Springer, 2011.

[20] S. Mohammad. #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255, Montréal, Canada, 2012. Association for Computational Linguistics.

[21] S. M. Mohammad, S. Kiritchenko, and X. Zhu. 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, June 2013.

[22] S. M. Mohammad and P. D. Turney. 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, 2010.

[23] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 2013.

[24] M. Naaman, J. Boase, and C.-H. Lai. Is It Really About Me?: Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192, New York, NY, 2010. ACM.

[25] K. Nishida, R. Banno, K. Fujimura, and T. Hoshide. Tweet Classification by Data Compression. In *Proceedings of the International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, pages 29–34. ACM, 2011.

[26] A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC*, 2010.

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

[28] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Philadelphia, PA, 2002.

[29] L. Pearl and M. Steyvers. Identifying Emotions, Intentions, and Attitudes in Text Using a Game with a Purpose. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California, 2010.

[30] R. Plutchik. A General Psychoevolutionary Theory of Emotion. *Emotion: Theory, research, and experience*, 1(3):3–33, 1980.

[31] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking "Gross Community Happiness" from Tweets. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, CSCW '12, pages 965–968, New York, NY, 2012. ACM.

[32] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM, 2010.

[33] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, 2009. ACM.

[34] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. 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*, SIGIR '10, pages 841–842, New York, NY, 2010. ACM.

[35] C. Strapparava and A. Valitutti. 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, 2004.

[36] K. Tsagkalidou, V. Koutsonikola, A. Vakali, and K. Kafetsios. Emotional Aware Clustering on Micro-blogging Sources. In *Proceedings of the Conference on Affective Computing and Intelligent Interaction*, pages 387–396, Memphis, TN, 2011.

[37] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*, 29(4):402–418, 2010.