# Emotion analysis on text mining for president prediction

Lang Gao

Computer & Information Science Department

Temple University

Philadelphia, PA, 19122

***Abstract*—President prediction is always a hot social topic as a president plays significant role in all aspects of the government agenda. It is therefore of great interest to know the next president in advance. By applying text mining on the most recent tweets gathered in April, we measured the public opinions on presidential candidates based on eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust. Compared to traditional text mining methods which categorize emotions as neutral, negative, and positive, our method provides more details and perspectives for the analysis of emotions and therefore yields a more accurate description of the emotion status. Our experiment demonstrates the feasibility of using text mining for president prediction.**

## I. Introduction

Applying opinion mining on twitter has been explored substantially and has promising applications such as stock market prediction [1], evaluation of movie success [2], etc. As a popular way of communication, tweets are often messages about the user's personal life, opinions on a certain topic, etc. They often convey pertinent information about the users' emotional states [3] and is therefore a good candidate for text mining. Through the use of tweet API, large volume of tweets on literally all possible topics can be collected, depending on the time scope of this collection process. After preprocessing, the data can be fed into the emotion analyzer to yield the overall emotion results for president prediction.

In this project, we proposed an emotion analysis framework that is applied on twitter for president prediction. The raw tweets data are preprocessed using the technique presented by Kumar and Sebastian [4] to give the resulting adjectives, adverbs, and verbs and a linear equation is used to find the emotion values in five categories, namely: happiness, anger, fear, sadness, and disgust. The overall sentiment scores are compared and the president is predicted based on the score of happiness.

This paper is organized as follows: Section 2 briefly analyze the background work of research on our topic, section 3 explains our sentiment analysis framework, section 4 presents the results of our analysis and corresponding discussions. Section 5 summarize the research undertaken and discuss the challenges, defects, and future work on twitter opinion mining

## II. Background Work

Traditional sentiment classifier on Twitter often determine sentiments for a document as positive, negative, and neutral. Alexander Park et al. [3] built a sentiment classifier based on a multinomial Naïve Bayes classifier which uses N-gram and POS-tags. It yielded the best result when compared with support vector machine and conditional random field.

EmpaTweet [5] provides an emotion corpus that is annotated with seven different emotions: anger, disgust, fear, joy, love, sadness, and surprise. By labeling a finer-grained set of emotions, they detected a greater range of emotion and ensured that their corpus has a greater lexical variety.

Akshi et al.[6] built an emotion analysis framework in which adjectives are analyzed across five different vectors: happiness, anger, sadness, fear, and disgust. However, they didn't address the situation where no adjective is present in the sentence.

Using Twitter text mining to predict president is still new, though researchers have used twitter for election prediction using twitter [7], but only the volume of twitter is used as the criteria, not the opinions or emotions behind the tweets.

## III. Emotional Analysis Framework:

The emotional analysis frameworks consists three parts: data collection, data preprocessing module and scoring module.

### A. Data collection

Publically available twitter dataset is collected through twitter's API, which is open to developers. Access can be granted if provide a valid pair of consumer key and consumer secret, as well as a pair of access token and access secret. In our project, the geo-location of the published tweets is specified as US, however it's not possible for us to limit the twitter users to American citizens only. Since twitter limits the access of tweets content to the past seven days, we collected tweets spread across eight days considering the time limit of this project. In the end we have 250,000 collected tweets from April, 2016, for each of the five candidates: Donald Trump, Hillary Clinton, John Kasich, Ted Cruz, and Bernie Sanders.

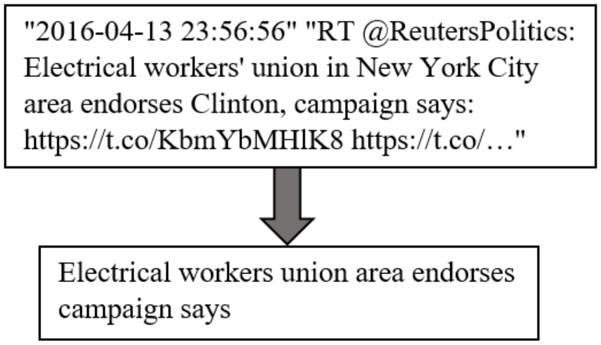
### B. Preprocessing for data cleaning

The raw twitter data collected has many undesirable symbols. We went through the following processes to clean the data:

1. Isolate “not” from words such as “hasn’t”, “don’t”, “haven’t”, “didn’t”, etc.
2. Remove hashtags, URL links, twitter user names, twitter special words such as “RT”, digits, punctuations, non-Unicode characters.
3. Tag the adjectives, nouns (proper nouns are excluded), adverbs, and verbs using the NL Processor Linguistic Parser.

An example on the effect of preprocessing is illustrated in graph I.

Graph I. Effect of preprocessing the twitter dataset.



"2016-04-13 23:56:56" "RT @ReutersPolitics: Electrical workers' union in New York City area endorses Clinton, campaign says: https://t.co/KbmYbMHlK8 https://t.co/…"

Electrical workers union area endorses campaign says

Table I. sample of words emotion value vector

| English word | Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust |
|---|---|---|---|---|---|---|---|---|
| accomplice | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| accomplish | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| accomplished | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| accomplishment | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| accord | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| accordance | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| accordion | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| account | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| accountability | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| accountable | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Only adjectives, nouns, adverbs, and verbs are remained after the preprocessing. The resultant words of each tweet are then feed into the scoring module to generate the scores measured as eight vectors.

### A. Scoring module

We used the NRC Word Association Lexicon [8] for the emotions. For each word, the lexicon includes eight vectors: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Versions are available for 20 languages, of which we only use the English version. A sample of the sentiment association is given in table I.

#### 1. Adjectives and nouns scoring

We can assign corresponding emotion values for all the nouns and adjectives by utilizing the above mentioned lexicon. The overall scores for each tweet is measured by adding the values of each vector for all the adjectives and nouns.

#### 2. Verbs and adverbs scoring

To get the strength of adverbs, we created a seed list of those that are either positive or negative. It’s then grown by searching for synonyms and antonyms from WordNet [9]. A sample is given in Table II. The strength of adverbs ranges from -1 to 1.

Table II. Adverb strength

| Adverb | Strength |
|---|---|
| completely | 1 |
| most | 0.9 |
| totally | 0.8 |
| extremely | 0.7 |
| too | 0.6 |
| very | 0.4 |
| pretty | 0.3 |
| more | 0.2 |
| much | 0.1 |
| any | -0.2 |
| quite | -0.3 |
| little | -0.4 |
| less | -0.6 |
| not | -0.8 |
| never | -0.9 |
| hardly | -1 |

For the adverbs that are presented in conjunction with verbs and adjectives, we multiply the emotion value of verbs and (1+sum of the strength of adverbs); for those adverbs that are isolated in a tweet, we still use the NRC Word Association Lexicon to obtain the related emotion values. If “not” is found associated with any other word, the orientation of the group of words is reversed by multiply “-1”.

The sum for each of the eight vectors are calculated for each tweet. The final emotion score for each vector of each candidate is found by taking the average over all the collected tweets.

## IV. Results and analysis

The result of the emotion analysis is given in table III. Since the emotion vectors for the tweets are rather sparsely distributed across the entire data set, all of their values are below 0.3. It’s not straightforward from the result to predict who has the best chance to become the president of the United States. In order to draw that conclusion, we have to give each vector reasonable weight, which relies on further physiological evaluation. Feelings are mixed for all the candidates. We can still find some interesting information. For example, people are most angry with Donald Trump, but meanwhile also most happy about him. Hillary Clinton tops the fear score but is also most trusted. If consider only anticipation and trust, she has the best change to be the next president.

| Name | Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust |
|---|---|---|---|---|---|---|---|---|
| Trump | 0.224 | 0.136 | 0.211 | 0.090 | 0.252 | 0.186 | 0.229 | 0.113 |
| Clinton | 0.167 | 0.210 | 0.190 | 0.186 | 0.202 | 0.174 | 0.132 | 0.227 |
| Cruz | 0.125 | 0.202 | 0.110 | 0.035 | 0.168 | 0.166 | 0.058 | 0.198 |
| Kasich | 0.201 | 0.025 | 0.177 | 0.169 | 0.172 | 0.111 | 0.220 | 0.205 |
| Sander | 0.187 | 0.086 | 0.204 | 0.177 | 0.168 | 0.180 | 0.210 | 0.177 |

Table III: Tweeter emotion analysis result for president prediction

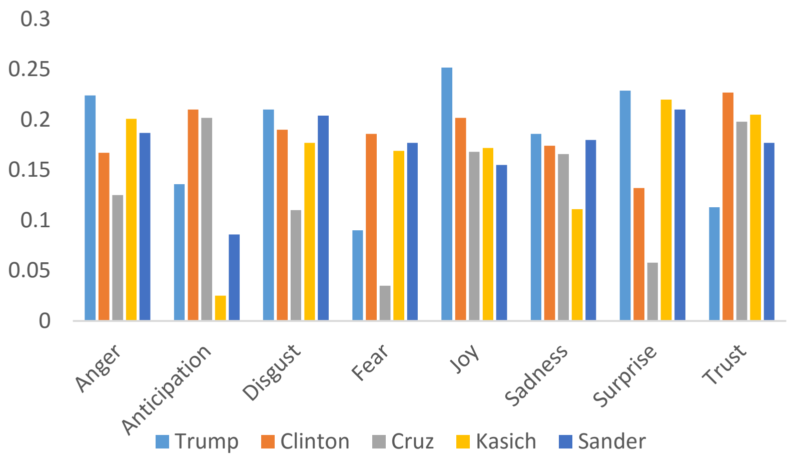


Figure II: Tweeter emotion analysis result for president

## V. Conclusion

In this project, we have applied text mining on Tweeter for president prediction for the upcoming election. Our work shows the feasibility of using emotional analysis on predicting result of some major events (in our case is the presidential election). Unlike conventional approaches, which consider the text subject as three vectors, namely, positive, negative, and neutral, our method provide a more detailed and in depth evaluation and the result illustrate the emotional state reflected in the text content from more perspectives. A more comprehensive understanding of the subject matter is therefore achieved. We also argue that even analyzing emotions from eight different perspectives is not enough to make the final prediction, as feelings are mixed towards the candidate and we don’t have a standard to analyze the combination of those different emotion vectors. To advance on this task, we need to come up with a criteria to assign different weights for each of the emotion feature. More data should be collected to increase the accuracy. The volume of the tweets also need be analyzed since it is different for different candidates and candidate who has large volume should get the credit. Moreover, a good understanding of the political environment is also important for successful prediction.

## VI. Acknowledgement

The author would like to specially thank Dr. Zoran Obradovic for his helpful guidance and valuable suggestions on this project.

**(5) References.**

1. Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.
2. Jain, Vasu. "Prediction of movie success using sentiment analysis of tweets." The International Journal of Soft Computing and Software Engineering 3.3 (2013): 308-313.
3. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREc*. Vol. 10. 2010.
4. Kumar, Akshi, and Teeja Mary Sebastian. "Sentiment analysis on twitter." *IJCSI International Journal of Computer Science Issues* 9.4 (2012): 372-373.
5. Roberts, Kirk, et al. "EmpaTweet: Annotating and Detecting Emotions on Twitter." *LREC*. 2012.
6. Kumar, Akshi, Prakhar Dogra, and Vikrant Dabas. "Emotion analysis of Twitter using opinion mining." *Contemporary Computing (IC3), 2015 Eighth International Conference on*. IEEE, 2015.
7. Shi, Lei, et al. "Predicting US primary elections with Twitter." *URL: http://snap. stanford. edu/social2012/papers/shi. pdf* (2012).
8. Mohammad, Saif M., and Peter D. Turney. "Crowdsourcing a word–emotion association lexicon." *Computational Intelligence* 29.3 (2013): 436-465.
9. “WordNet: A lexical database for English” The Trustees of Princeton University, 17 March, 2015. Web. 27 April, 2016.