Who Will Win? Donald Trump or Hillary Clinton
Forecast the Winner of Presidential Election
from Public Emotion in Twitter

by

ALSHEHRI Muhyi*1 and Iwao FUJINO*2

Abstract

Last November, all of the world concern about the presidential election of the United States of America and wonder who will win? Donald Trump or Hillary Clinton. Can we make an information system to forecast the winner? This is the purpose of our study. Because the voters post tweets entirely of their own accord, Twitter can mostly reflect the public emotion regarding each candidate. Without doubt, Twitter becomes the main medium of breaking news and people opinion directly than conventional mass media in recent year. Therefore in our study, in order to forecast who will win this election, we decide to choose Twitter as the original data source for measuring public emotion and we utilize geocoding web service to limit Twitter user only come from the United States. By counting the positive and negative emotional words for each candidate, we have succeed to forecast the result of this import election.

1. Introduction

Twitter is a social networking site that allows registered users to share tweets (i.e. messages) with their fans. It only allows a tweet of 140 characters maximum length. It became extremely gigantic system of exchanging messages between all people in the world. Twitter was created in 2006. Now it has 317 million monthly users [1]. Also it has 56.8 million active user in the United States up to 2016 [2] and that is about 21% of the population of adults. For the presidential Election in the United State of America in 2016, without doubt, Twitter is the main medium that people can express their personal emotions regarding each candidate. During recent years, a plenty of research has been done in the analysis of emotional polarity from social network such as twitter, Facebook and similar sites. Many researchers elaborate different techniques for handling this problem. These techniques can be classified into three categories: Manual techniques, semi-automated techniques and fully automated techniques. Das et al. [3] developed what is called domain-dependent lexicon that is considered as manual approach, whereas Hu et al. [4] introduced a semi-automated approach. On the other hand, Turney [5] presented a fully automated approach. The OpinionFinder

subjectivity lexicon was used by O'Connor et al. [6] for labeling the emotional polarity of tweets that were talking about the precedent president Barack Obama and compare quotidian emotional words scores with series of manually gathered approval ratings of Obama. In our research, we have chosen Twitter as the original material for measuring public emotion and we have used geocoding web service to limit Twitter user only come from the United States. By counting the positive and negative emotional words for each candidate coming from people, we have succeed to forecast the result of this important election. We will summarize our founding in the following points:

1. Our approach focus on a social media where the voters show opinion by themselves directly, not a conventional mass media. Because the voters post tweets entirely of their own accord, so Twitter can mostly reflect the public emotion about each candidate.

2. Our approach utilize geocoding web service to limit the tweets posted from an assigned country, state or other small area very precisely. Here for forecasting the presidential election, we gather tweets only from American people by limit Twitter user indicates his location of United States,

3. Our approach has good realizability, we can give prediction result day by day, certainly before the Election Day, even on a relatively low performance computer.

As a result of our research, even with this simple polarity determination, we found significant result that is quite different from the opinion of some highly skilled and professional journalists. Many of them believed that Hillary Clinton would win the election, whereas our prediction results that were obtained from our system showed that Donald Trump who is going to win and this is certified by the announcement of the presidential election results.

*1 Graduate School of Information and Telecommunication Engineering, Course of Information and Telecommunication Engineering, Master's Program

*2 School of Information and Telecommunication Engineering, Department of Communication and Network Engineering, Professor

2. Methodology

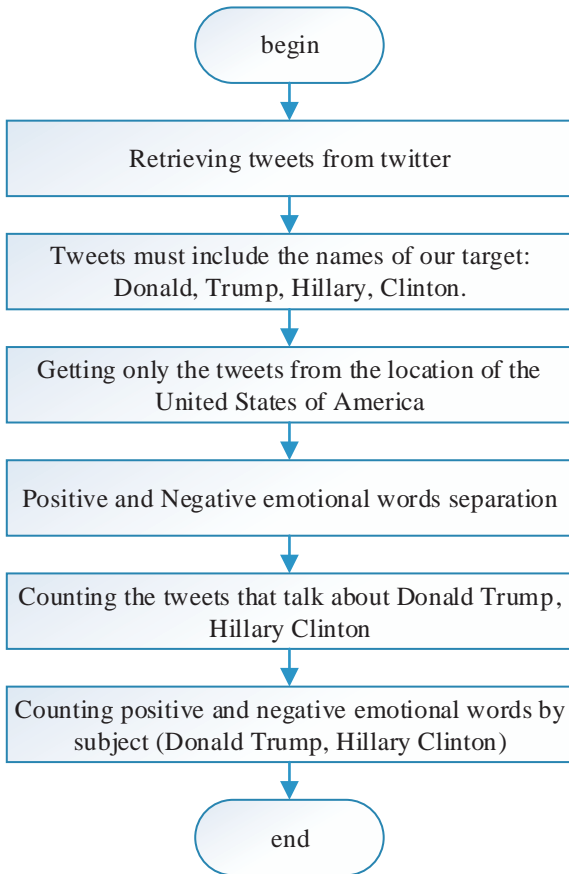The proposed method is described with a flow chart as shown in Fig. 1.

begin

Retrieving tweets from twitter

Tweets must include the names of our target: Donald, Trump, Hillary, Clinton.

Getting only the tweets from the location of the United States of America

Positive and Negative emotional words separation

Counting the tweets that talk about Donald Trump, Hillary Clinton

Counting positive and negative emotional words by subject (Donald Trump, Hillary Clinton)

end

Fig. 1: Flow Chart of Procedures to Count positive and negative emotional words by subject

We collect Tweets from Twitter Streaming Data and save these data to CouchDB according to its date. For utilizing geocoding web service we use a software library called Geopy[+1]. Also for preparing positive and negative emotional words list, we download subjectivity lexicon from MPQA[+2] web site.

3. Experiment Results

3.1 *Collected Tweets and Sampled Tweets*

We collected Tweets from Oct 26, 2016 to Nov 17, 2016, whereas the Election Day is Nov 9, 2016. Then we counted these tweets day by day and the result is shown in Fig. 2 with the label noted by "The number of tweets". Meanwhile,

because of the restriction of geocoding web service, getting the location in all collected tweets cannot be done unless we register in the Geopy for having the code to get the username and password. However, it is very expensive for us. Therefore, we had no choice but to use the program within the range of its free access. The maximum of free access is about 3500 times per day. For this reason, we could only use a sampled portion of the collected data as shown in Fig. 2, with the label noted as "Number of sampled tweets".
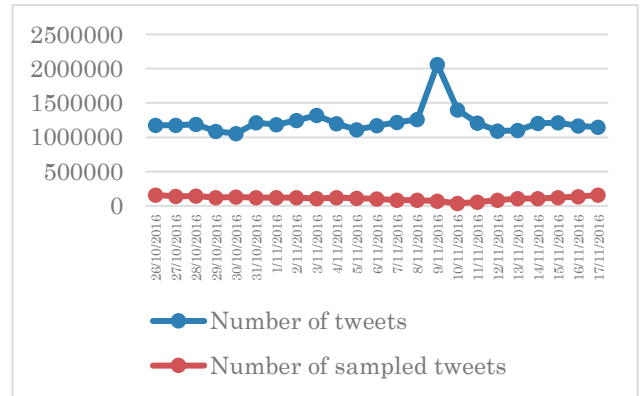

Fig. 2: Number of real tweets vs number of sampled tweets

3.2 *Tweets Regarding Donald Trump and Hillary Clinton*

Fig. 3 shows the number of tweets about Donald Trump and Hillary Clinton in the sampled tweets during the 23 days. The activity on twitter showed that the number of tweets talking about Hillary Clinton was greater than that of Donald Trump before the day of election. After the Election Day, tweets about Donald increased tremendously w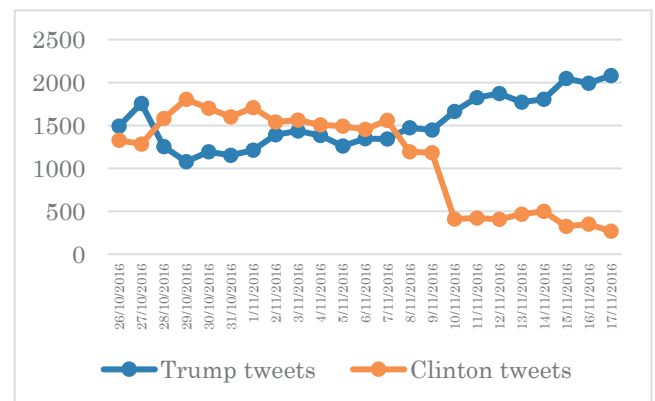hereas those of Hillary Clinton were dramatically vanishing. However, this does not necessarily mean who will win or lose the election. We will see in the following the expectation of who win and who lose.

Fig. 3: Donald Trump and Hillary Clinton Tweets in 23 days

+1 https://pypi.python.org/pypi/geopy

+2 http://mpqa.cs.pitt.edu/

3.3 *Donald Trump and Hillary Clinton Positive Emotional Words*

From Fig. 4, we can see that the emotional positive words about Donald Trump are quite higher than those of Hillary Clinton in most days before the Election Day. Also, on this day, tweets concerning Donald Trump tends more positively than those towards Hillary Clinton. Moreover, during the days after the Election Day, Donald Trump got more mentions on twitter with positive emotional words than Hillary Clinton. The only period in which Hillary Clinton tends more positively than words towards Donald Trump is between Oct 29-Nov 1 and Nov 7.
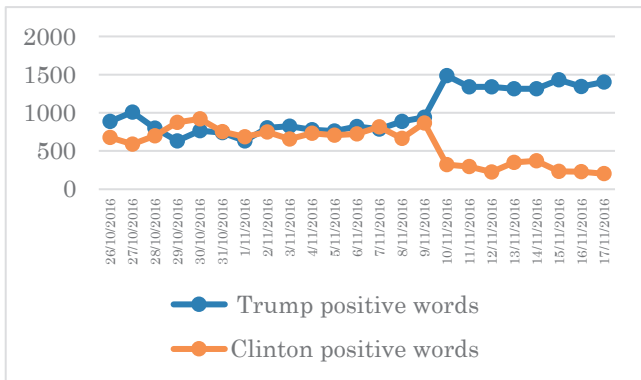

Fig. 4: Donald Trump and Hillary Clinton Positive Emotional Words

Table 1 shows the total average positive words per candidate in pre-election days. We can see that Donald Trump has more positive words than Hillary Clinton.

Table 1: Average of Positive Words

	Average of positive words
Donald Trump	805.4
Hillary Clinton	742.1

3.4 *Donald Trump and Hillary Clinton Negative Emotional Words*

Next from Fig. 5, we can clearly see that the emotional negative words about Donald Trump are a little bit lesser than those of Hillary Clinton before the Election Day especially between Nov 6 and the Day of Election. On the other hand Hillary Clinton tends to be more negatively emotional than words towards Trump during the period that starts from Oct 28 to Nov 2. Also, on the Election Day and after that, words concerning Donald Trump tends more negatively than those towards Hillary Clinton. Even though it seems not good about him but it means that Trump got more mentions on twitter with positive and negative emotional words than Hillary Clinton as shown in Fig. 4 and Fig. 5.
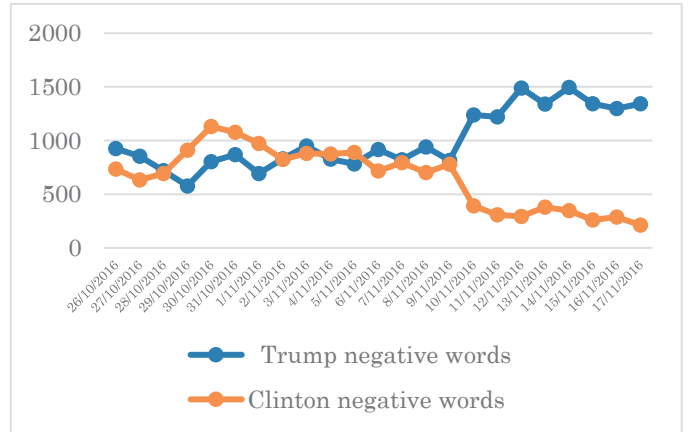

Fig. 5: Donald Trump and Hillary Clinton Negative Emotional Words

Table 2 shows the total average negative words per candidate in pre-election days. We can see that Donald Trump has less negative words than Hillary Clinton.

Table 2: Average of Negative Words

	Average of negative words
Donald Trump	821.3
Hillary Clinton	840.8

From these results shown in Table 1 and Table 2, we can forecast that Donald Trump will win this presidential election before the day at election. This is a very different result comparing with that from conventional mass media.

4. Analysis

4.1 *Pre-Election Days*

We have calculated the emotional positive and negative words per candidate as shown in Fig. 6 in pre-election days and we can clearly see that when it comes to Hillary Clinton the percentage of emotional negative words are greater than that of positive words.
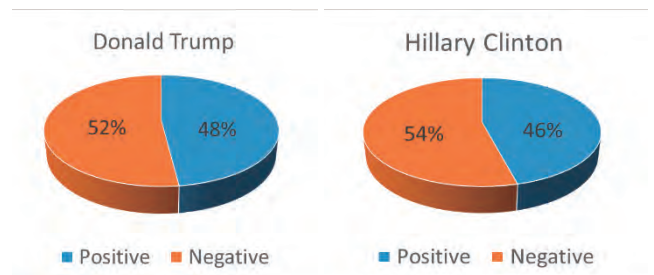

Fig. 6: Percentage of Positive and Negative Words Per Candidate in Pre-Election Days

The same thing can be said for Donald Trump as well, however, it is less compared to that of Clinton. In order to simplify this result in a more clear vision, Fig. 7 shows that the percentage of emotional positive words are greater when it comes to Donald Trump whereas negative words are greater when it comes to Hillary Clinton. From these graphs, we can expect that Donald Trump has greater chance to win the election against Hillary Clinton. We will further measure the same thing after the Election Day and see whether this expectation was correct or not.
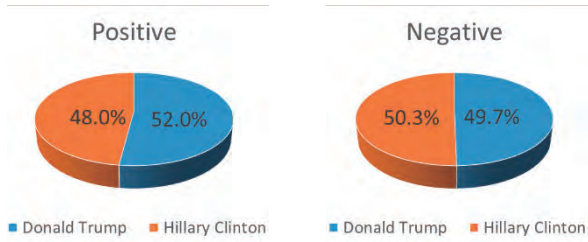

Fig. 7: Percentage of Positive and Negative Words for Both Candidates in Pre-Election Days

4.2 *Post-Election Days*

In order to make sure that the emotional trends on Donald Trump and Hillary Clinton in Twitter are stable and give continuous support to our predicted result, we further made some analyses on the tweets of post-election days. As it clearly can be seen in Fig. 8 and Fig. 9 that Donald Trump got more attention as he won the election and the talking about Hillary Clinton started to diminish in both aspects either negative or positive words in all the tweets we have examined.
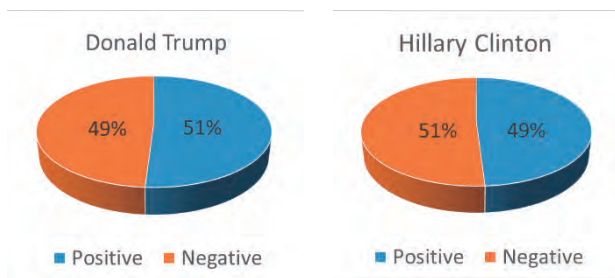

Fig. 8: Percentage of Positive and Negative Words per Candidate in Post-Election Days

If we compare the positivity percentage concerning Trump pre-election and post-election, we can see that the emotional positive talk about him increased from 48% to 51% whereas the emotional positive talk about Hillary Clinton increased from 46% to 49% as shown in Fig. 6 and Fig. 8.
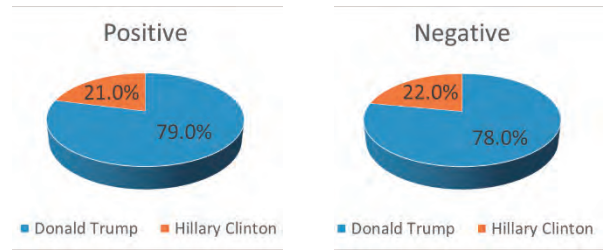

Fig. 9: Percentage of Positive and Negative Words for Both Candidates in Post-Election Days

5. Conclusions

As a result of our study, by choosing Twitter as the original data source for measuring public emotion and using a geolocation web service, we can get tweets about each candidate only from the Unite State. Then by counting the positive and negative emotional words for each candidate in tweets, we have succeed to forecast the winner of the presidential election of the United States of America, while most of conventional mass media have given wrong result.

References

[1] Isaac Mike, Ember Sydney, For Election Day Influence, Twitter Ruled Social Media, The New York Times, November 8, 2016. Retrieved November 20, 2016.

[2] Number of Twitter users in the United States from 2014 to 2020, https://www.statista.com/statistics/232818/active-us-twitter-user-growth/. Retrieved June 10, 2017.

[3] S. Das and M. Chan. Extracting market sentiment from stock message boards. Asia Pacific Finance Association, 2001, 2001. Retrieved May 12, 2016.

[4] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177, New York, NY, USA, 2004. Retrieved May 12, 2016.

[5] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 417–424, 2002. Retrieved May 12, 2016.

[6] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge and Noah A. Smith, From tweets to polls: Linking text sentiment to public opinion time series, In Proceedings of the International AAAI Conference on Weblogs and Social Media, 2010. Retrieved May 12, 2016.