TOWSON UNIVERSITY

OFFICE OF GRADUATE STUDIES

MEASURING APPROVAL RATING FOR ELECTION IN TWITTER

by

Jongsung You

A thesis

Presented to the faculty of

Towson University

in partial fulfillment

of the requirements for the degree

Master of Science

Department of the Computer and Information Sciences

Towson University

Towson, Maryland 21252

May, 2013

TOWSON UNIVERSITY
OFFICE OF GRADUATE STUDIES

THESIS COMMITTEE APPROVAL FORM

Student's Name

Jongsung You

Chair, Thesis Committee Yanggon Kim

Signature Typed name

Member Sungchul Hong

Signature Typed name

Member YUANQIONG WANG

Signature Typed name

Member

Signature Typed name

Note: Please attach a description of the affiliation and credentials of any non-Towson University members of the Committee, and the members' *curriculum vitae*.

Approved by

Graduate Program Director Date May 10. 2013

Signature

Department Chairperson 5/10/2013

Signature Date

Dean of Graduate Studies May 22, 2013

Janet V DeLany

Signature Date

ABSTRACT

With the advances in Social networks, many fields use Social networks service to increase user. The most popular Social networks service is Twitter. Companies use Twitter to get information about film advertisement, product advertisement, and approval ratings for such thing like the president. In the previous research, we searched influence nodes related to political fields using the Data Gathering Tool and gave weight to the influence nodes. Users express opinion's using tweets on Twitter that is both positive and negative. Many researchers study positive and negative tweet to use for a survey. So we measure presidential approval rating using influence weight and positive and negative weight and compare the measurement results and real approval ratings. In conclusion, we confirm whether Twitter approval ratings reflect real approval ratings or not.

TABLES OF CONTENT

Table Page

LIST OF FIGURES

Figure Page

1 Introduction

With the advent of the web 2.0, many users express their own opinions on the Internet with rapid Internet growth. These opinions are used as the public opinion analysis of the people about government policy. The company confirms the information about the product or service and product reliability [8]. Recently, the users of Social networks service have increased. On the Internet, the formation of relationships between users having common matter of concern is supported. It is the service which it applies in order to act the various activities including the personal relationship management, information and contents share. Based on the acquaintance relationship formed in this way. With the increasing popularity of many online social network sites, such as Facebook, Twitter, Blogger, LinkedIn, and MySpace, a massive amount of data has become available.

Twitter is the real-time information network used to contact the latest news of the user’s fields of interests, ideas, opinions, and recent news. Tweets are publicly visible by default, but senders can restrict message delivery to just their followers who are connected to the other users. Users can tweet via the Twitter website. The tweets were initially set to a 140-character limit for compatibility with Social networks messaging service. By using this tweet, its own news is updated and the sentiment for the service or specified product is expressed and the information is shared to other users. [1][5][6][7] In June 2010, Twitter experienced rapid growth, and about 65 million tweets were posted daily. After two years, March 2011, that was about 140 million tweets posted daily. [2] The users are influenced by Twitter because of the rapid growth and the application of data mining techniques to online social media benefits many groups, such as market researchers, psychologists, sociologists, businesses, and politicians showing fascinating insights into human behavior, marketing, business, or political views [19][20].

The consumer can use the sentiment analysis so that he/she can investigate the product or service before buying. The marketing staff can use Twitter in order to research the opinion of this company and product and analyze the customer's satisfaction. In addition, the company uses Twitter in order to collect the definite feedback about the problems of the new product opened to the public. Even in the field of politics, Social networks service like Twitter or Facebook is used in order to collect user's opinions. The Orange's Digital Election Report of BBC systematically analyzes the political leverage of the digital media about the general election in Great Britain. It answered about 24% among the young voters are uninterested in politics write text about the election or politics for the period of general election through Social networks service including Twitter and Facebook. [14] In fact, during the period of the general election in Great Britain, it was exposed to be actually similar to the election results to the measured public opinion in Twitter in the political information site called the Twittermaster(tweetminster.com).

However, it is difficult to discover useful information from social data without automated information processing because of three main characteristics of social media data sets: the data is large, noisy, and dynamic. In order to overcome these challenges of social media, data mining techniques can be used by data seekers to discover a diversity of perspectives that would not be possible otherwise. Data mining techniques are widely used to handle large sets of data and to discover new knowledge and useful information in a data set that is not readily obtainable and not always easily detectable.

The sentiment research of classification analyzes the opinion extracts and the opinion mining including positive and negative activity. Based on natural language processing and machine learning research to determine the opinion of positive, negative[9][10] and based on a statistical analysis, there is the research that it grasps the score of the users about the product and assessment of the feature unit from the feature vocabulary frequency of the product. [11][12][13]

There is much research that finds the person where there is influence in Twitter. Users will trust Influencers in Twitter among the opinion of the Influencer and user. So, it has the interests in the research we find the Influencer. This research makes a study whether the approval rating of the presidential candidate of Twitter reflects to the approval rating of the real presidential candidate or not.

We use the method analysis of two kinds of polarity analysis and influencers. Firstly, I analyzed the polarity. For example, the negative and positive tweets of Twitter are analyzed and then the quantity of the polarity ware analyzed. Second, we find the influencer. We think the tweet of the Influencer more important than the tweet of the other users. For example, negative ten tweets value of the user and negative one tweet value of the Influencer can be same.

By using these two methods, our research assigns the other weight for each influencer, and the weight is assigned to tweets of the negative and positive. By using this weight, we analyzed the approval rating on Twitter. And then, we made a research whether this approval rating can reflect the real approval rating or not. If it reflects, we made a study how it is reflected.

The remainder of this paper is constructed as follows: In Section 2, the related works that have been done so far are summarized. Section 3 introduces approach for calculating predictable weight and explains details of the algorithm. Section 4 presents the results of data gathering and compares two approval ratings calculated by the algorithm. The last part, Section 5 concludes the work by summarizing this paper.

2 Related work

2.1 Twitter Polarity Classification with Label Propagation

There are the various methods with the research related to the polarity classification. The tweet includes the information, spoken language and clipped words. Michael Speriosu and other researchers [17] approach in order that the word expresses the negative, positive, and emoticon. Michael Speriosu and other researchers approaches various ways through the label propagation using the Modified Adsorption Algorithm. The Label Propagation was used the Jnuto Label Propagation. A Toolkit, Toolkit's implementation of MAD, uses the Modified Adsorption algorithm. Michael Speriosu and other researchers researched whether express Twitter flow graph or not. The researchers extract the emoticon using the training data set about "garden hose" from Twitter API. Michael Speriosu and other researchers extracted among 6.265,345 tweets in which the emoticon is included. Out of extracted tweets 5.156,277 of the positive emoticon and 1.109,068 negative emoticons. In the same way, the data set of the Stanford Twitter Sentiment (STS), [15] Obama-McCain Debate (OMD) [16] and Health Care Reform (HCR) were used in order to prove the accuracy of the polarity classification. Michael Speriosu and other researchers used the Amazon Mechanical Turk annotates on the positive, negative and neutral tweets, during the presidential debate on September 26, 2008 between Barack Obama and John McCain, Michael Speriosu and other researchers extracted the label of the positive, negative about the gold standard. The results showed 705 negative gold tweets and 1,192 positive gold tweets were extracted among the total 1,898 tweets. HCR is the new data set about the Health Care Reform. Michael Speriosu and other researchers applied a label propagation method where sentiment labels were propagated from a small set of nodes seeded with some initial label information throughout the graph. Michael Speriosu and other researchers claimed that their label propagation method outperforms MaxEnt trained from noisy labels and obtained an accuracy of 84.7% on the subset of Twitter sentiment test set from STS. Table shows accuracy percentages.

Classifier	STS
Ramdom	50.0
LEXRATIO	72.1
EMOMAXENT	83.1

LPROP(Follower-edges, Maxent-seed)
LPROP(All-edges, Lexicon-seed)
LPROP(Feature-edges, Noisy-seed)
LPROP(All-edges, Noisy-seed)
	83.1
	70.0
	84.7
	84.7

Table 1: Per-tweet accuracy percentages. The models and parameters were developed while tracking performance on STS results was obtained from a single, blind run.[17]

Michael Speriosu and other researchers constructed a graph that has some of the microblogging features, such as hashtags and emoticons, together with users, tweets, word unigrams and bigrams, as its nodes which are connected based on the link existence among them Users are connected to tweets Michael Speriosu and other researchers created; tweets are connected to word unigrams that Michael Speriosu and other researchers contain.

Consequently, the Label Propagation uniting the several knowledge sources and noisily supervised label propagation algorithm of the accuracy was more improved than the other method. Michael Speriosu and other researchers did not find overall gains from using the following graph as implemented here.

2.2 Measuring User Influence in Twitter: The Million Follower Fallacy

Social media has become extremely popular. Many companies spend 1 billion dollars in the Social Media for marketing. This is used for the political campaign, contents sharing, and advertisement of products. So, the advertiser finds the influence user. However, people misunderstand that the influence that increases public size and the number of flowers were high influencer. So, in Meeyoung Cha's paper[21], by using Twitter, the User Influence was measured. As to the reason, Twitter is one of the most popular social networks. Because it is connected between the users with the flow, the user can find easily which interest in any user and topic. In 2009, 54 million users, 2 billion follow links, and 1.7 billion tweets were collected. These three activities represent the different types of influence of a person:

1. Indegree influence: the number of followers of a user, directly indicates the size of the audience for that user.

2. Retweet influence, which we measure through the number of retweets containing one’s name, indicates the ability of that user to generate content with pass-along value.

3. Mention influence, which we measure through the number of mentions containing one’s name, indicates the ability of that user to engage others in a conversation.

In other words, Indegree is the number of people who follow a user. “retweet” mean the number of times others “forward” a user's tweet (RT), and “mention” mean the number of times others mention a user's name (@username). The relative ranks of a user across three measures using Spearman’s rank correlations. Table 2 is shown in order to investigate how the three measures correlate; we compared the relative influence ranks of all 6 million users.

Table 2: Spearman’s rank correlation coefficients. [21]

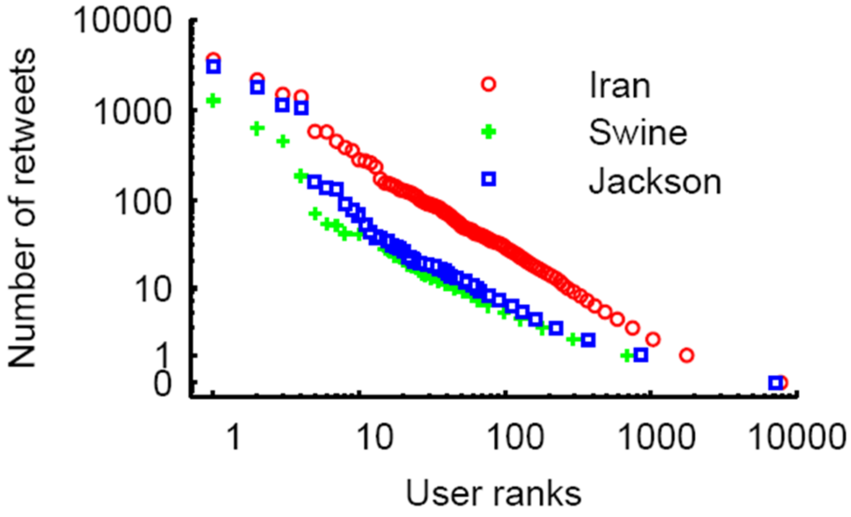
Correlation	All	TOP 10%	Top 1%
Indegree vs Retweets	0.549	0.122	0.109
Indegree vs Mentions	0.638	0.286	0.309
Retweets vs Mentions	0.580	0.638	0.605

Consequently, the Indegree generally correlates with retweets and mentions. In 2009, the most popular kind of the three topics selected ware Jackson, Swine, Iran. As shown in the table, there are the user, tweet, and audience of concern in the Iran, Swine, and Jackson topics. Table 3 is shown in.

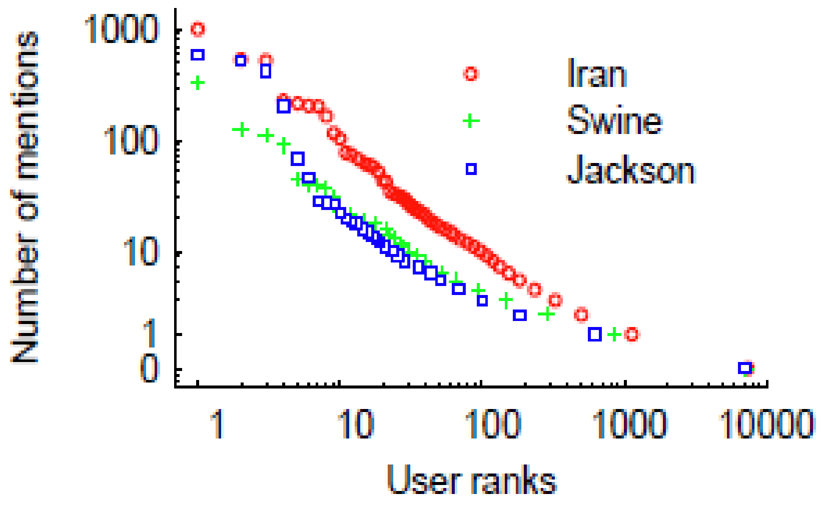
Table 3: Summary information of the three major topics events studied [21]

Topic	Users	Tweets	Audience
Iran	302,130	1,482,038	22,177,836
Swine	239,329	495,825	20,977,793
Jackson	610,213	1,418,356	23,550,211

As follows, the user link for a given topic is distributed.



(a) Retweet influence ranks



(b) Mention influence ranks

Figure 1: Distribution of user ranks for a given topic. [21]

Depending on the high user rank, it can confirm that there are lots of retweets or mentions. In Meeyoung Cha's paper, Meeyoung Cha and other researchers would like to analyze results so Indegree can measure the user's popularity. However, it cannot think that there is the important influence. Retweet refers to the value of the tweet and the mention refers to the user is value.

2.3 Politics, Elections and Data

There is the research to utilize politics in the field as the social influence of the smart device and Social networks service gradually increases. The political community in the election is no exception. The case in which the political party or candidate utilizes Social networks service for the election campaign was noticeably increased. In the election, Social networks service was used actively.

The paper of JamesFosco and other researchers [22] analyzes the most important topics about the US presidential election in 2012. By using TOPSY (www.topsy.com) service, the political tweets were collected. After classifying the collected Twitter by subjects, It made the rank list. The method of measurement uses the comment, the mentions and retweets in order to trace the influence. The tweets having the keyword by subjects were collected and classified. From October 16th until November 6th in 2012, over 800,000 tweets and ware analyzed. By using the power law distribution, JamesFosco and other researchers analyzed whether the users have concern about any kind of topics. The ranking of topics was the economy, Foreign Policy, Health Care, Abortion, Same-Sex Marriage, Immigration, Education, and Gun Rights. This ranking results the total tweets of the results of the Obama and Romney election. If this result is compared by the base with other Poll result, for the economy, health care, and Foreign Policy, three kinds of topics coincide. The negative and positive meaning is included in the collected tweets. The polarity was analyzed according to each topic, and according to the flow of the time, James Fosco and other researchers analyzed the tweet value. The location of the peak of the tweet about each candidate ware confirmed to coincide with the actual topic of the candidate. The graph shows that actual issues have an effect on Twitter.

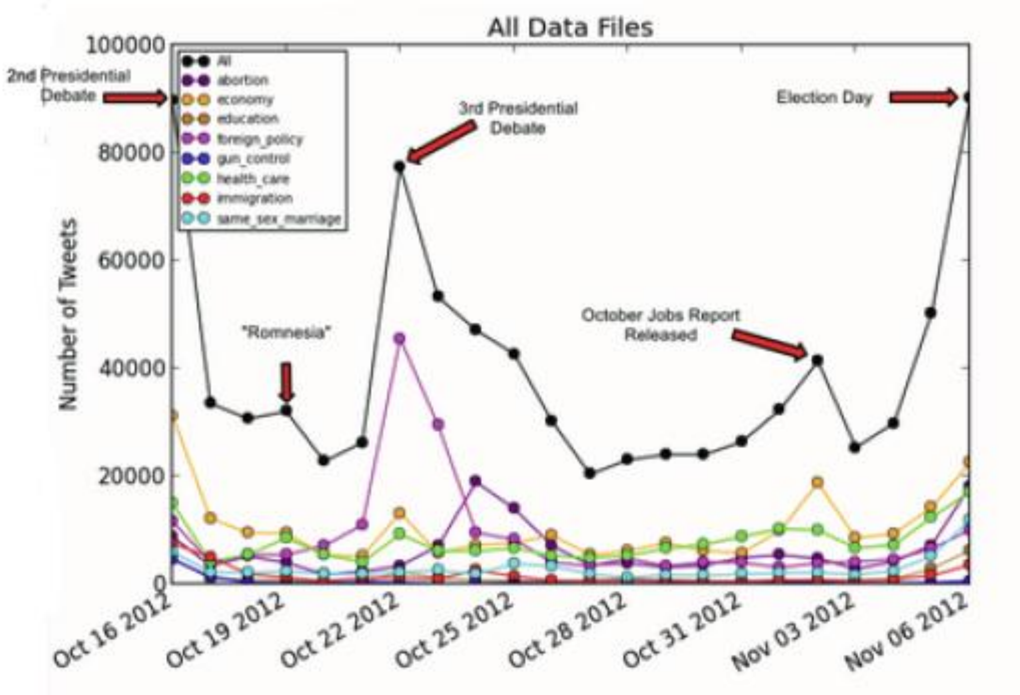


Figure 2: The number of tweets collected in all of our data plotted over time.

The cases of other countries, by using Twitter, Tjong Kim Sang and Johan Bos [25] studied the prediction of the Dutch Senate Election. Tjong Kim Sang and Johan Bos counted the tweets mentioning the political part for the Dutch Senate Election in 2011, and it tested whether the expectation of the election result was right or not. They extracted the tweets about the Dutch Senate Election in 2011. As to the result of polls and Twitter, there were differences. As to the reason, a person can make many tweets on Twitter, and there are the voters looks like the old people, who have hard a time using the computer. So, the predictions of the election results of the vote have the difference. Tjong Kim Sang and Johan Bos[25] tested two normalization steps for Twitter data. First, Tjong Kim Sang and Johan Bos[25] removed all tweets that mentioned more than one party name. Next, Tjong Kim Sang and Johan Bos[25] kept only the first tweet of each user. Finally Tjong Kim Sang and Johan Bos[25] combined both steps. The results can be found in Table4 and Table5. The offset of three methods proved that approach with the normalization is effective than the approach without the normalization.

Table 4: Population weights per party resulting from dividing the percentage of the predicted poll seats [25]

Party	**One party Per tweet**	**One tweet Per user**	**Both Twitter**
PVV	22	17	19
VVD	12	13	13
CDA	12	12	12
PvdA	8	8	8
SP	6	8	7
GL	6	7	7
D66	5	5	5
CU	1	2	2
PvdD	1	1	1
SGP	1	1	0
50+	0	0	0
OSF	1	1	1
offset	29	22	25

Table 5 : Twitter seat prediction for the 2 March 2011 Dutch Senate elections compared with the actual resultsand the predictions of two polling companies of 1 March 2011[25]

Party	**One party Per tweet**	**One tweet Per user**	**Both Twitter**
PVV	811	0.49	13
VVD	552	0.68	13
CDA	521	0.70	12
PvdA	330	0.69	7
SP	314	0.90	9
GL	322	0.81	9
D66	207	0.94	6
CU	104	0.67	2
PvdD	63	1.00	2
SGP	39	0.86	1
50+	17	0.93	0
OSF	-	-	1
offset			23

Next, Tjong Kim Sang and Johan Bos [25] determined the sentiments of the tweets. Tjong Kim Sang and Johan Bos [25] used these 1,333 tweets with unanimous class assignment for computing sentiment scores per party. Tjong Kim Sang and Johan Bos [25] computed weights per party by dividing the number of nonnegative tweets per party by the associated total number of tweets. For example, there were 42 negative tweets and 89 nonnegative. Twitter in a weight of 89/(42+89) = 0.68. The resulting party weights can be found in Table 5 .Consequently, the poll prediction decreased from 25 to 23 from Table 4 and Table 5.

We confirmed whether the collected results in Twitter were reflected in the results of the US presidential election in 2012 or not. Many researchers have studied how use of Twitter by politicians and citizens relates to results of public opinion polls and elections. Tumasjan and other researchers [23] argue that Twitter message content reflects the offline political landscape, thus potentially predicting actual election results. In a Tumasjan's case study, numbers of tweeted messages were observed to closely match ranking by share of the vote in election results, and nearly approximated results of traditional election polling. We analyzed the polarity at each tweet for the high accuracy and the weights are calculated with the polarity. Also, we distinguished the Influencer. According to the size of the influence, the weight is given. These two weights are combined. It confirms whether or not, the two results are reflected to the results of the actual election.

3 Methods

There is numerous research used in the calculation of an approval rating using Twitter. We were going to introduce an algorithm to calculate influence and polarity weight to find relations between activities in Twitter and an approval rating. The two factors for calculating predictable weight are activity weight from the Data Gathering Tool and Positive/Negative Points. In the previous research [24], we presented a java-based data gathering tool with the Seed Analysis module. Firstly, it continuously and automatically collects data from Twitter. Secondly, it allows us to collect keyword-related data. Thirdly, it more efficiently collects data from only qualified nodes. Fourthly, it stores collected data into database for analysis. Fifthly, it calculates activity weight of each qualified nodes. Finally, it supports intuitive user interface to interact with users.

3.1 Approach Twitter Data Collecting Tool with activity weight

This data gathering tool uses the seed analysis module. The seed analysis module includes the Initial Node Selecting Algorithm and the Node’s Activity Calculating Algorithm. Figure 3 shows architecture of the Twitter data.

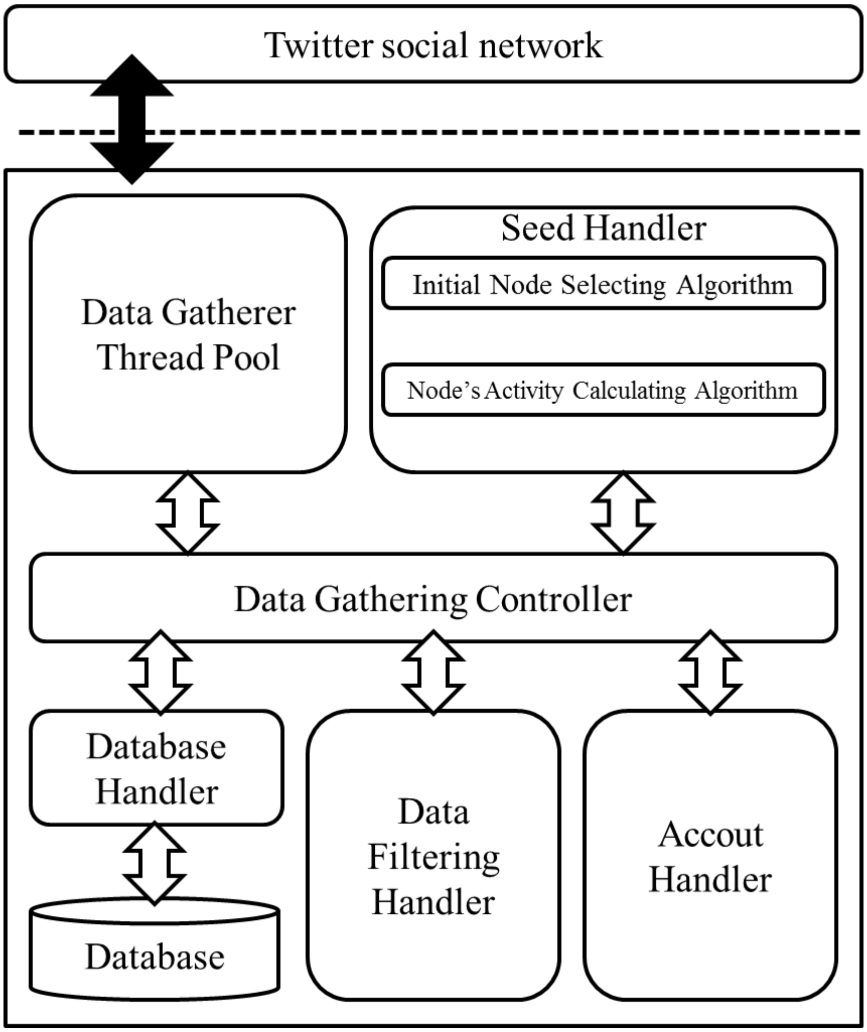

Twitter social network
Data Gatherer Thread Pool
Seed Handler
Initial Node Selecting Algorithm
Node's Activity Calculating Algorithm
Data Gathering Controller
Database Handler
Database
Data Filtering Handler
Accout Handler

Figure.3. Architecture of the Twitter data gathering tool with seed analysis module

We used the algorithm for selecting influential nodes. As previously mentioned, an interest in finding influential nodes in social networks is increasing. Social influence analysis is aimed to either demonstrate the existence of social influence or to quantify the strength of the influence.

We were collecting data from only qualified nodes. This goal can be achieved by giving activity weight to each node and checking if the node has enough activity weight before collecting tweets from the nodes. Figure 4 shows the data gathering process from selecting an initial node through storing tweets into the database.

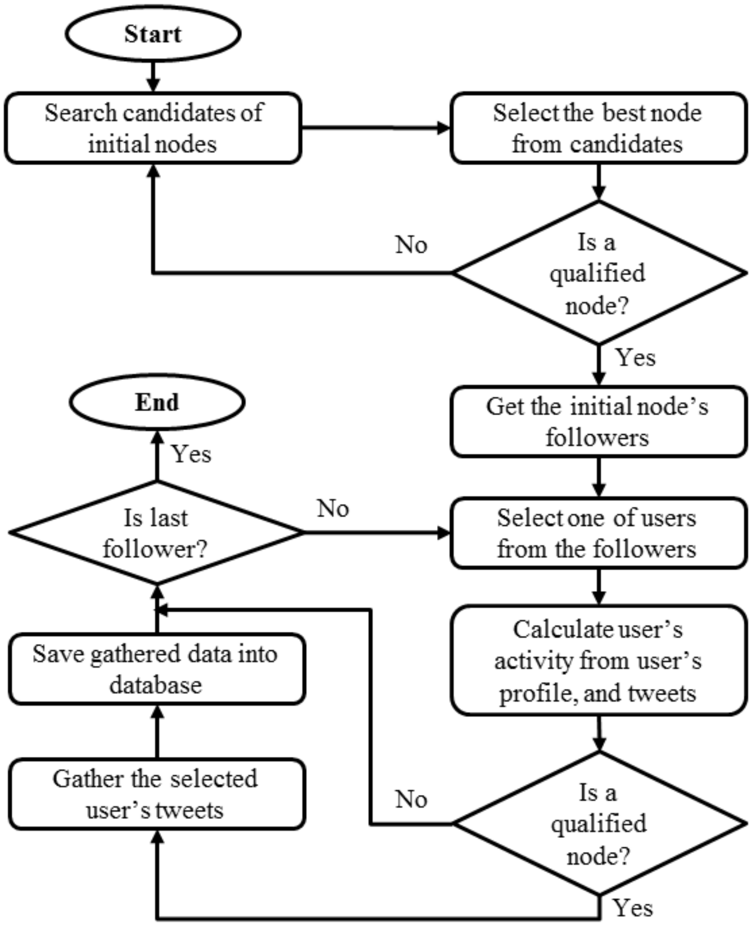

Start
Search candidates of initial nodes
Select the best node from candidates
Is a qualified node?
No
Yes
Get the initial node's followers
Select one of users from the followers
Calculate user's activity from user's profile, and tweets
Is a qualified node?
No
Yes
Gather the selected user's tweets
Save gathered data into database
Is last follower?
No
Yes
End

Figure. 4: Flow chart of data gathering process by qualified nodes[24]

At the beginning of the data gathering process, a list of candidates for an initial node is organized based on their most recent tweet, which includes a certain keyword at least once. When the list of candidates is built, an algorithm calculates the activity weight of each candidate, and the list is organized by the number of followers of each candidate node. The first node in the list that has the highest number of followers and is a qualified node will be the initial node. If the first node in the list is not the qualified node, then check the next available node to see whether or not it is qualified. This process of finding a qualified node is iterated until an initial node is selected. The tool generates a list of the initial node's follower information, such as each follower's unique id, language, number of followers, or number of friends. Then, each follower's activity weight is calculated. Only tweets from qualified users are collected, until there is no more follower information on the list. Figure 5 shows the simple algorithm used to find a node, which has more followers than others.

Notation: A is a set of nodes such that recently posted their tweets about a certain keyword. B is a pointer indicating the best initial nodes in the set A.

```
Set B=A[0].
Loop I from 1 to A.size – 1.
    If B.no_of_followers < A[I].no_of_followers, then
        Set B= A[I].
    End If.
End Loop.
End of Algorithm
```

Figure.5. Algorithm of selecting initial node from a list of candidates[24]

If the selected node is not a qualified node, the tool removes the node from the list and runs the algorithm to find a suitable node again. In a user's profile, there are properties to be considered as factors of the user’s activities in Twitter, such as number of followers, number of friends, number of keyword-related tweets, date tweeted, and favorite count. Among the user’s properties, we use the number of followers, the number of keyword-related tweets, and dates tweeted as main factors for calculating user’s activity.

$$\text{Activity weight} = \frac{\text{the number of tweets containing the keyword}}{\text{the number of tweet}} \quad (1)$$

Formula 1: Activity weight

Notation: T is a set of tweets such that recently posted by a user within 30 days from search date and time. K is a string variable containing a keyword. W is a float variable containing user’s activity value calculated by this algorithm. M indicates the number of tweets and N implies the number of tweets containing the keyword W.

```
Set M=T.size.
Set N=0.
Loop I from 0 to M
    If the tweet ,T[I], contains the keyword K, then
        Set N=N+1.
    End If.
End Loop.
    If M is not 0, then.
        Set W=N/M.
    End If.
End of Algorithm.
```

Figure 6. Algorithm of calculating user’s activity weight[24]

3.2 Positive Negative

We collected data using the Twitter Data Collecting Tool. The keyword-related tweets of influential nodes collect and classify tweets of the data by negative tweets and positive tweets in the gathering tweet list. We extract a positive and negative word from politic-related tweets and make lists based on the OpenFolder. We extract 2000 positive words and 100 negative words from politic-related tweets. If one tweet has two positive words and one negative word, we regard it as a positive tweet. If it has an equal number of positive words and negative words in the tweet, we regard it as a natural tweet. Neutral tweets were removed from being counted. If a tweet includes positive keywords in the list, it is a positive tweet, and if a tweet includes negative keywords, it is a negative tweet. Table 6 shows the keywords for each positive and negative tweet.

Table. 6. The keywords for each positive and negative.

Category	Keywords
Positive words	promises, vote, help, believe, hope, excited, accuse, wonder, incredible ,imperial president ,triumph, shame ,love, accept, want, confirms , elected, wish, celebrates, enjoy ,praise, funny, proud, pledges, interest, excellent, approve, good, well, best, great, amazing, nice, pretty, cool
Negative words	wasteful, hate, never, lie, not, inexperienced, liar, stumble, demolishes, disappointed, repeal, worst, fault, blunders, unelected, fraud, traumatic, oppose, upset, offend, cheated, abandoning, usurps, sabotages, unlikely, drops out, protesting ,ignore ,betray, impeachment, ,impeach, contradicts, disgusting , destroy, article, failure, unconstitutional, reached, waste, terrible, destroyed,

	trouble, deny, failed, ‘t

Polarity categorization also calculates the same way as the activity weight. If a tweet has positive keywords, each tweet adds weight. If a tweet has negative keywords, each tweet subtracts weight. We aggregated the total positive and negative weight and used the formula (2) to calculate the polarity weight.

$$\text{Polarity weight} = \frac{(\text{Positive weight} + \text{negative weight})}{\text{number of keyword related tweets}} \quad (2)$$

Formula 2: Polarity weight

Figure 7 shows an algorithm that calculates the polarity weight of each date.

Notation: T is a set of tweets such that recently posted by a user within 30 days from search date and time. NK is negative words in Table 6. PK is positive words in Table 6. W is polarity weight calculated by this algorithm. D is number of tweets of each date. M indicates the number of tweets and P and N implies the number of tweets containing the positive and negative.

```
Set M=T.size.
Set P=0.
Set N=0.
Loop I from 0 to M
  Loop J from 0 to D
        If the tweet ,T[J], contains the keyword NK, then
           Set N=N+1.
        End If.
        If the tweet ,T[J], contains the keyword PK, then
           Set P=P+1.
        End If.
  End Loop.
End Loop.
       If M is not 0, then.
                Set W=P-N/D.
       End If.
End of Algorithm.
```

Figure 7. Algorithm of calculating polarity weight[24]

3.3 Define an algorithm to combine all factors together.

We calculated the Twitter approval rating using polarity weight and Activity weight. The figure is expressed as a percentage. We then used the formula (3) to calculate the Twitter approval rating.

$$\frac{\text{Polarity weight} + \text{Activity weight}}{\text{Total weight}} = \text{Twitter approval rating} \quad (3)$$

Formula 3: Twitter approval rating

Figure 8 shows architecture of the Twitter approval ratings of Twitter using Formula 3. First, the tool gathers tweets from Twitter’s API and calculates the weight of the activity node. Second, each activity node has a tweet that includes positive and negative words. So we calculated the polarity weight using Formula 2. Finally, we calculated the Twitter approval rating using Formula 3.

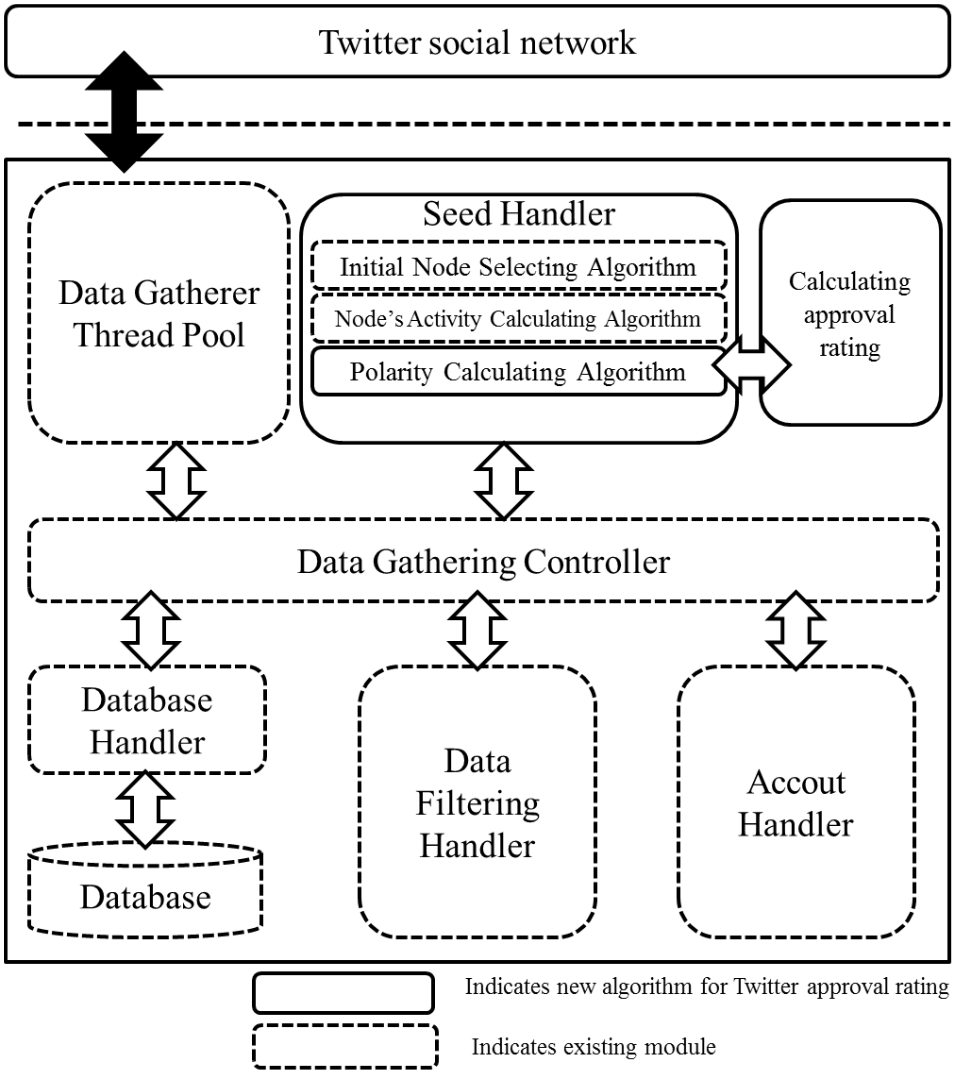

Twitter social network
Data Gatherer Thread Pool
Seed Handler
Initial Node Selecting Algorithm
Node's Activity Calculating Algorithm
Polarity Calculating Algorithm
Calculating approval rating
Data Gathering Controller
Database Handler
Database
Data Filtering Handler
Accout Handler
Indicates new algorithm for Twitter approval rating
Indicates existing module

Figure 8: Architecture of the Twitter data gathering tool with approval rating module. [24]

4 Experiment

In this section, we present an analysis of building a Twitter dataset with the Twitter data gathering tool, using our influential node selecting algorithm on the real Twitter network. We prove whether an approval rating of Twitter data can reflect a real approval rating or not.

There are three presidential debates until the presidential election. Each debate has topics and an open event on an average of one time a week. The U.S. Presidential debate is important; voters show interest in a debate because it will help them decide who to vote for. Almost voters watch the presidential debate on television. Few voters actually attend the debates. Table 7 shows the exact date and time of each debate. We made Table 7 to illustrate all of the election-related dates and times.

Table 7: Data and Time of the Presidential Election 2012

Event Name	**Topic**	**Debate date**	**Time**
First presidential debate (domestic policy)	Domestic policy	October 3, 2012	9:00-10:30 p.m. Eastern Time
Second presidential debate (town hall format)	Town meeting format including foreign and domestic policy	October 16, 2012	9:00-10:30 p.m. Eastern Time
Third presidential debate (foreign policy)	Foreign policy	October 22, 2012	9:00-10:30 p.m. Eastern Time

United States presidential election, 2012		November 6, 2012	

4.1 Political Twitter data with the Data Gathering Tool

Based on the date and time of each event illustrated in Table 7, we gathered tweets from Twitter. We collected the total number of tweets, which was 1,429 and divided the Twitter data by each candidate. Table 8 shows the number of tweets gathered from the Twitter Data Collecting Tool by each candidate. We gathered tweets from Table 8 and divided it into three categories: Obama, Romney, and Other. Obama data is the number of tweets related to Barack Obama. Romney data is the number of tweets related to Mitt Romney. Other is the number of tweets besides these two candidates. Obama data gathered 70 political tweets from September 26, 2012 to October 09, 2012; 34 political tweets from October 10, 2012 to October 16, 2012; 59 political tweets from October 17, 2012 to October 22, 2012; and 106 political tweets from October 23, 2012 to November 06, 2012. Romney data gathered 73 political tweets from September 26, 2012 to October 09, 2012; 30 political tweets from October 10, 2012 to October 16, 2012; 33 political tweets from October 17, 2012 to October 22, 2012; and 56 political tweets from October 23, 2012 to November 06, 2012.

Table 8: Number of gathered political tweet during 09/26/2012 - 12/13/2012

Date	Whole data	Obama data	Romney data	Other

09/26/2012 - 10/09/2012	307	70	73	164
2012/10/10 - 2012/10/16	182	34	30	118
2012/10/17 - 2012/10/22	190	59	33	98
2012/10/23 - 2012/11/06	504	106	56	342
Total	1183	269	192	722

Figure 9 shows the number of tweets gathered from the Twitter Data Collecting Tool by each candidate; this is the statistical data in a graphic form. We confirm the fact that an increase of political tweets during a presidential debate than other dates. So we considered the number of tweets is related to an approval rating through the result. Blue express whole data, red express Obama data, and green express Romney data. The Y of the graph is the number of tweets, and the X of the graph is the date.

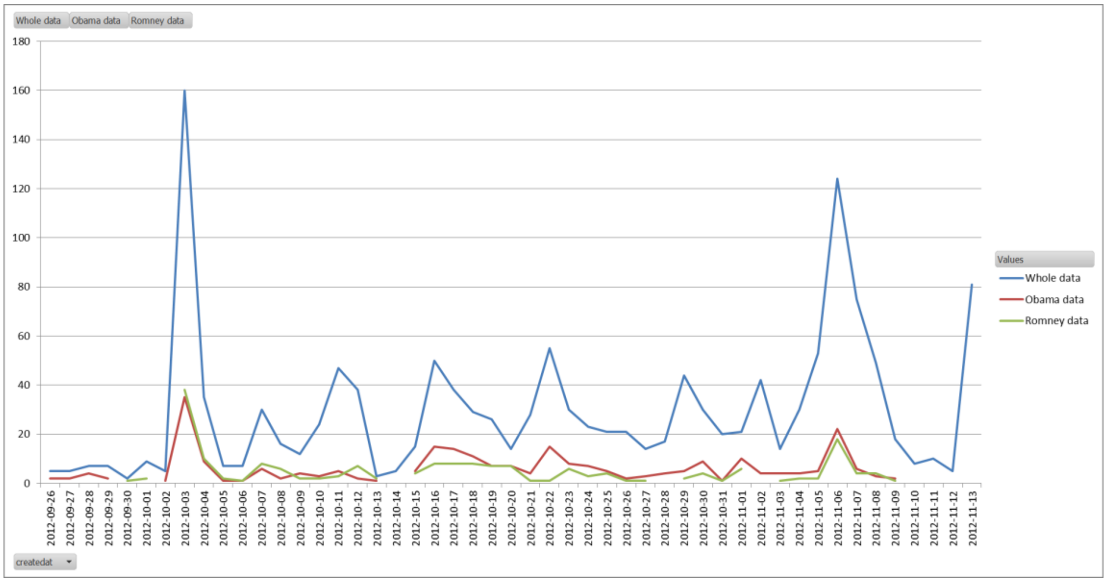


Figure 9: Graph of gathered political tweet during 2012/09/26 ~ 2012/12/13

4.2 Activity weight of influence node

Table 9 shows the weight of an activity node. We considered high influence nodes, which have the most followers and friends related to politics, and then calculated activity weight using formula (1). Table 9 is in order of high influence node from the Twitter Data Collecting Tool. User name is Twitter ID, number of followers, which means how many users follow the user, and number of following, which means how many other people the user follows. If it has many political tweets, we consider influence user and grant high weight. For example, there are 2 tweets related to politics in 13,196 tweets and there are 15 tweets related to politics in 18 tweets. We compared the two cases and considered that a node having 15 tweets is an influenced node related to politics.

Table 9: Activity weight of each activity node

NO	User name	Activity weight	Number of followers	Number of following
1	Sbo****	0.80	19	84
2	Tim****	0.42	318	1297
3	BRB****	0.30	49	206
4	1sa****	0.28	103	208
5	ITN****	0.26	105	234
6	San****	0.25	154	397
29	Mat****	0.01	5142	490
30	gab****	0.01	13196	9633
31	lao****	0.01	283	1084

4.3 Polarity classification

We classified tweets related to Obama as either positive or negative using polarity keywords in Table 6. If there are no positive and negative tweets, they are separated by using a neutral tweet. We classified Romney tweets in the same way. Table 10 shows the number of two candidate tweets including positive and negative tweets. The number of politic-related tweets was 18% positive tweet, 10% negative tweets, and anything else as a neutral tweet.

Neutral tweets are the largest in number and the next highest type is the positive tweet.

Table 10: Number of candidate positive and negative tweets

Event Name	Number of tweets			
	Total	Positive	Negative	Neutral
Obama	284	53	34	197
Romney	214	35	17	162
Total	498	88	51	359

Figure 10 shows that statistical data in a graphic form. Blue is the number of positive tweets, red is the number of negative tweets, and green is the number of neutral tweets.

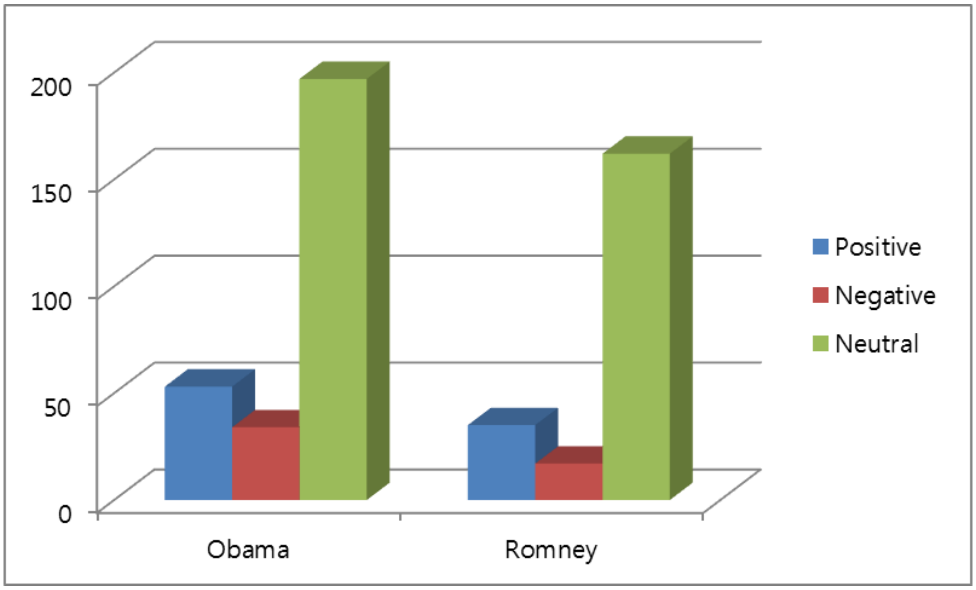


Figure 10: Number of positive and negative tweets

We calculated polarity weight using only positive and negative, and calculated weight using formula (2). There is sometimes a case when users mention positive tweets and mention negative tweets. So we calculated polarity weight that subtracts negative from positive. The point is divided by number of tweet of the day. Each user had a different point by date because total tweets of each day are different. Table 11 shows the weight of some users.

Table 11: Weight of candidate positive and negative tweets

Date	User name	Activity weight	Number of Positive	Number of Negative
2012/09/26 - 2012/10/09	Hap****	0.004	1	
	Tim****	-0.004		1
	BRB****	0.008	2	
2012/10/23 - 2012/11/06	sal****	-0.005		1
	Jgr****	0.005	1	
	Hap****	0.015	3	

4.4 Twitter approval rating

We calculated the total point, aggregating influence weight and polarity weight. Influence nodes had several tweets. The tweet may be a positive tweet or not.

Users tweet on Twitter more than once a day. We then added up the positive tweets and negative tweets of the user. The polarity weight of each influence node can be calculated using formula (2) and aggregate the result and influence weight. Next, add the result of the same date, we calculated the result by date. The result is each candidate point of Twitter. Figure11 shows the Twitter point of Obama by date. Figure12 shows Twitter point of Romney by date. We consider that influence voter in Twitter tweets many positive tweets during the presidential debate. Candidate Romney received the highest number of points during the first presidential debate and during the next presidential election. Obama received more points than average on Twitter during other presidential debates. Black expresses the average and blue is the candidate point.

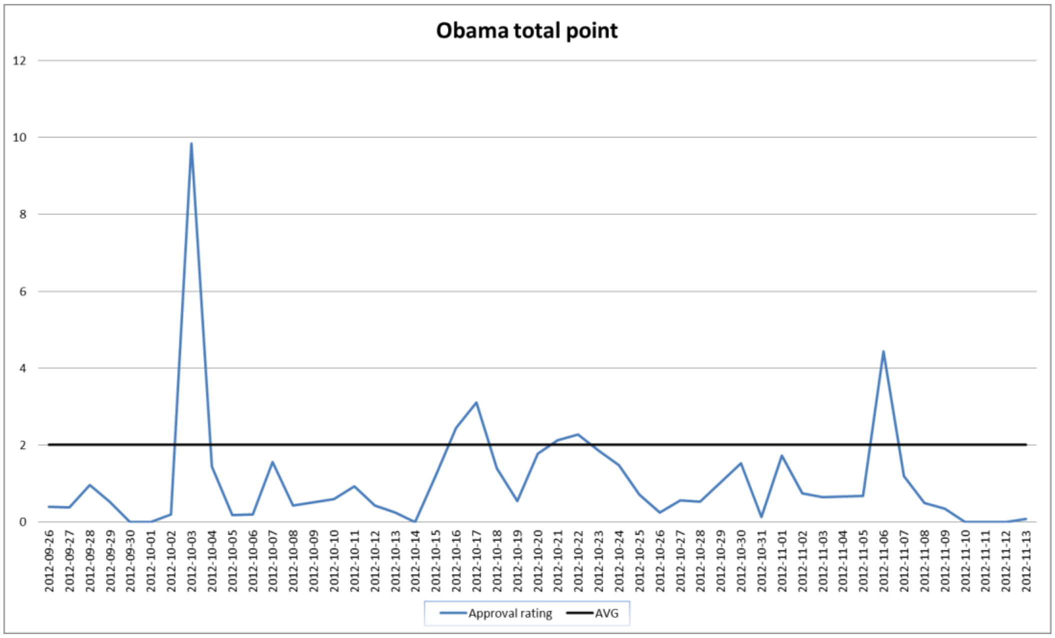


Figure 11: Twitter point of Obama

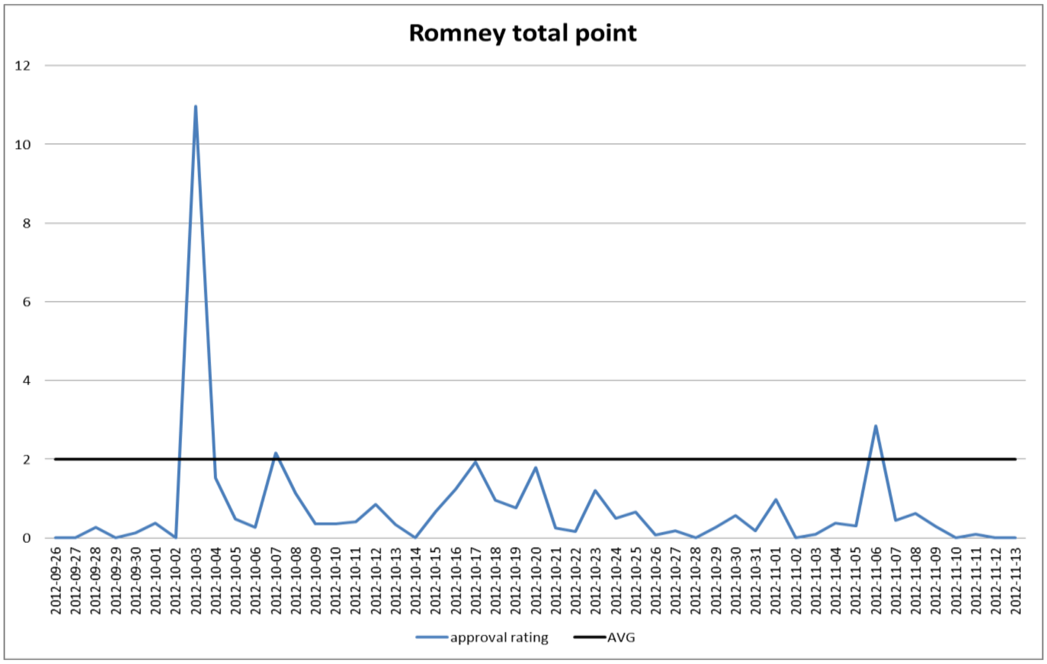


Figure12: Twitter points of Romney.

Figure 13 shows the results compared with two Twitter points given above. We considered which candidates' voters supported during each presidential debate. Romney received more points than Obama during the first presidential debate. During the second presidential debate and the third presidential debate, Obama received more points than Romney. Obama got higher points than Romney on Election Day. We confirmed that the graph result was similar to the real approval rating.

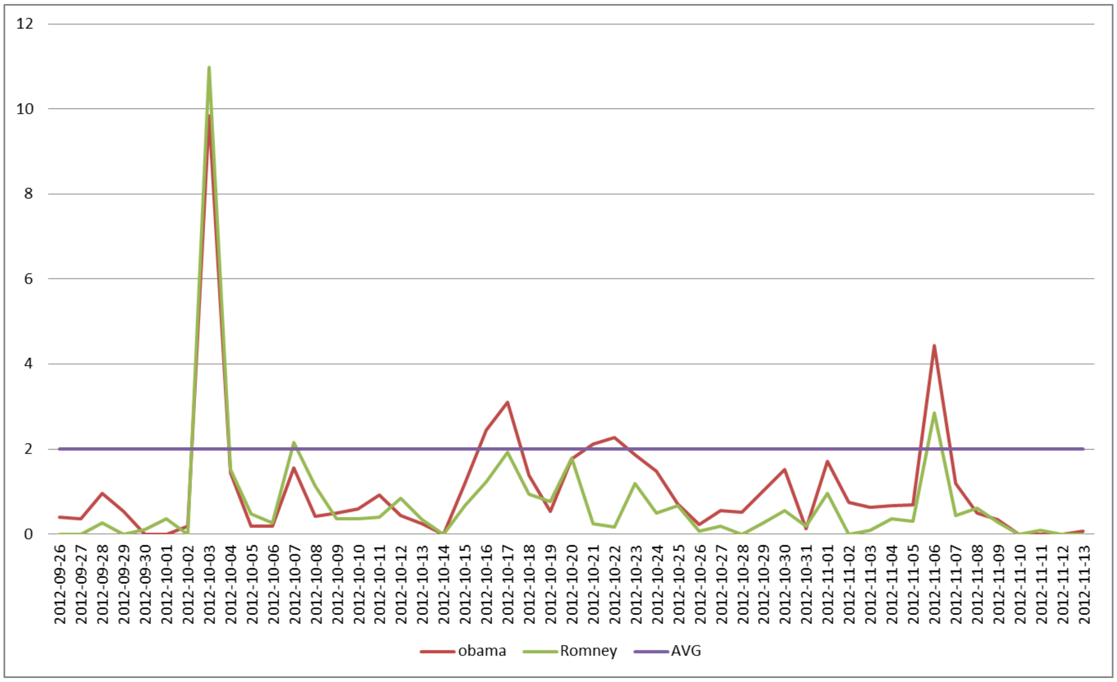


Figure 13: Twitter total point of both candidates

Table 12 shows the results of the Twitter point to calculate active weight and polarity weight and the results were analyzed in percentage terms to compare the real approval rating. We consider that the point was Twitter approval rating.

Table 12 Twitter approval rating

Event Name	Date	Obama (%)	Romney (%)	Both (%)
First presidential debate (domestic policy)	2012/09/26 - 2012/10/09	39.82%	43.12%	17.06%
Second presidential debate (town hall format)	2012/10/10 - 2012/10/16	45.78%	39.37%	14.85%

Third presidential debate (foreign policy)	2012/10/17 - 2012/10/22	42.98%	41.99%	15.02%
United States presidential election 2012	2012/10/23 - 2012/11/06	53.19%	46.81%	

4.5 Real data about the Presidential Election 2012

Table 13 shows data from The CNN Polling Center[1]. It displays data from polling organizations that use CNN-approved polling methodology. It only represents the views of people who watched the debate. There are three different types of viewers in this poll. The first presidential debate consisted of 430 American adults, who are 37% Democratic and 33% Republican. The second presidential debate is a total of 457 American adults who are 33% Democratic and 33% Republican. The third presidential debate showed 48% of watchers who believed Barack Obama would win and 40% of viewers who thought Mitt Romney would win. The polls were conducted after the presidential debates.

Table 13: Data gathering result by CNN

Event Name	Debate date	Obama (%)	Romney (%)	Both (%)	Watchers
First presidential debate (domestic policy)	October 3, 2012	25%	67%	8%	430 Democratic : 37% Republican : 33%

[1] http://www.cnn.com/POLITICS/pollingcenter/index.ht

Second presidential debate (town hall format)	October 16, 2012	46%	39%	12%	457 Democratic : 33% Republican : 33%
Third presidential debate (foreign policy)	October 22, 2012	48%	40%	5%	Debate watchers
United States presidential election, 2012	November 6, 2012	51.1% (65,907,213)	47.2% (60,931,767)		

4.6 Compare approval rating

Table 14 shows the results compared with Twitter approval rating and real approval rating. This result can be confirmed that the results will be the same through the comparison. Voters showed more support for Romney than Obama during the first presidential debate. They showed more support to Obama from the second presidential debate to the presidential election in 2012. Average error was 19.35% in the first presidential debate. The average error was 0.29% in the second presidential debate. The average error was 3.50% in the third presidential debate, while the average error was 1.19% in the presidential election in 2012. The total average error was 5%-6%. So we decided to reflect real approval rating using Twitter approval rating. Figure 14 and Figure 15 show that we made a bar graph to compare CNN approval rating result and Twitter approval rating result. Blue is Obama approval rating and red is Romney approval rating.

Table 14: Compare with Twitter approval rating and Debate result

Event Name	CNN approval rating result		Twitter approval rating result	
	Obama	Romney	Obama	Romney
First presidential debate	25%	67%	39.82%	43.12%
Second presidential debate	46%	39%	45.78%	39.37%
Third presidential debate	48%	40%	42.98%	41.99%
United States presidential election, 2012	51%	47%	53.19%	46.81%

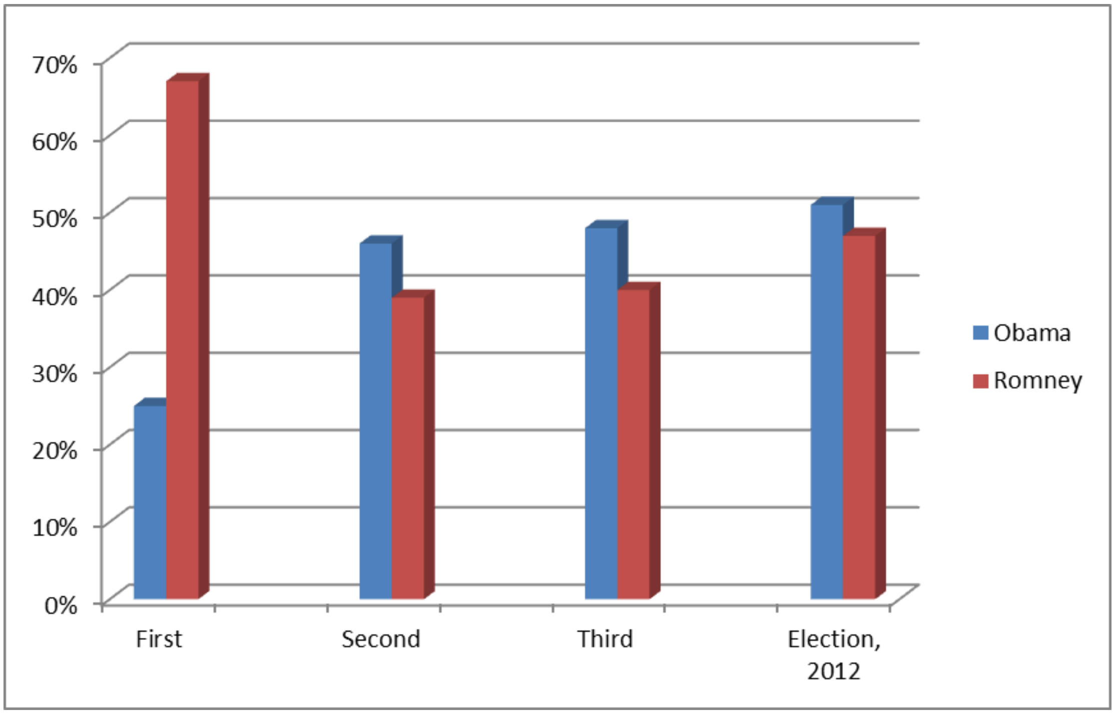


Figure 14: CNN approval rating result

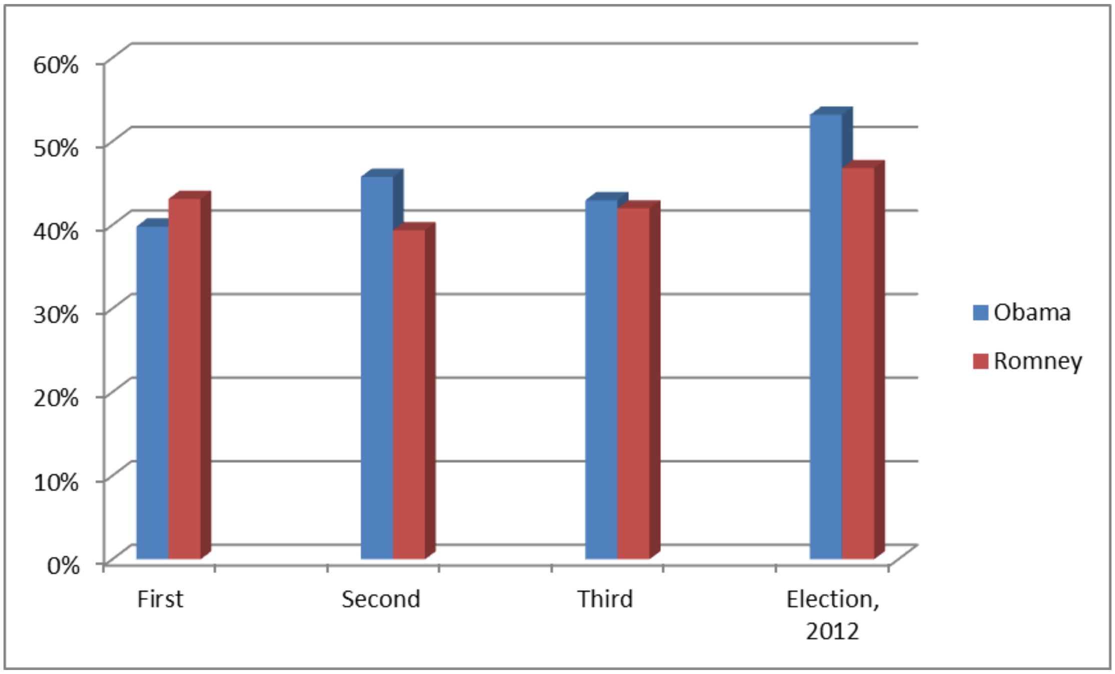


Figure 15: Twitter approval rating result

5 Conclusion

Recently, the users of Social networks service have increased. The most popular Social networks service is Twitter. Researchers collected opinions of user easily because users follow user interest. They researched to positive and negative tweets and to find influence users. Researchers use Twitter in the politics area to survey who to vote for. Researchers want to find out political influence user in Twitter, and to divide with positive and negative tweets included meaning that has the emotions of people. So we studied these research related polarity classification and finding influence user. We develop the java-based data gathering tool to collect user massages (tweets). We calculated influence of the collected tweets and gave high weight to user who has many political mentions of collected tweets. We divided into positive and negative tweet. There were 53 positive and 34 negative tweets in 284 Obama related politics and 35 positive and 17 negative tweets in 284 Romney related politics. We calculated a polarity weight to subtract negative weight from positive weight. We added the both weight and convert into Twitter approval rating. CNN surveyed approval rating during the first presidential debate, the second presidential debate and the third presidential debate. We compared Twitter approval rating and real approval rating from CNN. The result compared with Twitter approval rating and real approval rating from CNN, an average error was 5%~6%. We prove that Twitter approval rating reflect real approval rating.

Future work on using our result, Twitter result is as applicable to other field. Company advertise new product and surveyed. If they use our research method, Companies collect customer opinion easily and they can measure a hot item. Other field, we give weight to tweet in Twitter. We can know a major event in Twitter.

References

[1] L. Barbosa and J. Feng, "Robust sentiment detection on Twitter from biased and noisy data", In Proceedings of the 23rd International Confere-nce on Computational Linguistics, pp.36-44, 2010.

[2] Beaumont, Claudine, "Twitter Users Send 50 Million Tweets Per Day – Almost 600 Tweets Are Sent Every Second Through the Microblogging Site, According to Its Own Metrics", The Daily Telegraph (London). Retrieved February 7, 2011.

[3] Kristian Nicole Smith," Social Media and Political Campaigns," Chancellor's Honors Program, May 2011

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classication using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing

[5] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, 2004, pp.168-177.

[6] Xiaowen Ding, Bing Liu and Philip S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", WSDM'08, 2008, pp.231-239.

[7] W.Y.Kim, J.S. Ryu, K.I.Kim, U.M.Kim, "A Method for Opinion Mining of Product Reviews using Association Rules", ICIS, 2009, pp.270-274.

[8] Scaffidi, C., Bierhoff, K., Chang, E.,Felker, M., Ng, H., and Jin, C., "Red Opal: Product-Feature Scoring from Reviews," In Proceedings of the 8th ACM conference on Electronic Commerce

[9] Esuli, A. and Sebastiani, F., "Determining Term Subjectivity and Term Orientation for Opinion Mining," In Proceedings of 11th conference of the European chapter of the Association for Computational Linguistics

[10] Jin, W., Ho, H., and Srihari, R., "OpinionMiner : a novel machine learning system for web opinion mining and extraction", In Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining, 2009.

[11] Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T., “Mining Product Reputations on the Web”, In Proceedings of the 8th SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[12] Liu, B., Hu, M., and Cheng, J., “Opinion observer : analyzing and comparing opinions on the Web”, In Proceedings of the 14th international conference on World Wide Web, 2005

[13] Ding, X., Liu, B., and Yu, P. S., “A holistic lexicon-based approach to opinion mining”, In Proceedings of the international conference on Web search and web datamining, 2008.

[14] Painter, Anthony, “A Litte More Conversation, A Little More Action: Orange’s Digital Election Analysis”, www.Orange.co.uk. 2010.

[15] Alec Go, Richa Bhayani, and Lei Huang,”Twitter sentiment classification using distant supervision”, Stanford University, 2009.

[16] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill, “Tweet the debates: understanding community annotation of uncollected sources”, In Proceedings of the first SIGMM workshop on Social media, pages 3–10, 2009.

[17] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. “Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph”, In Proceedings of the EMNLP 2011 Workshop on Unsupervised Learning in NLP, pages 53-63, 2011.

[18] Partha Talukdar and Koby Crammer,“New regularized algorithms for transductive learning”, In Wray Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor, editors, Machine Learning and Knowledge Discovery in Databases, volume 5782, pages 442–457. Springer Berlin / Heidelberg, 2009.

[19] J. C. Cortizo, F. M. Carrero, J. M. Gomez, B. Monsalve, and P. Puertas, “Introduction to mining social media”, In F. M. Carrero, J. M. Gomez, B. Monsalve, P. Puertas, and J. C. a. Cortizo, editors, Proceedings of the 1st International Workshop on Mining Social Media, pages 1–3, 2009.

[20] I. King, J. Li, and K. T. Chan,” A brief survey of computational approaches in social computing”, In IJCNN’09: Proceedings of the 2009 international joint conference on Neural Networks, pages 2699–2706, Piscataway, NJ, USA, 2009. IEEE Press.

[21] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy”, Proc. International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010

[22] James Fosco, Charlie Fierro,”Twitter Political Influence - Presidential Election 2012”. Stanford University, 2012

[23] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. ,” Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Retrieved from

[24] Changhyun Byun , Hyeoncheol Lee , Jongsung You , Yanggon Kim “Dynamic Seed Analysis in a Social Network for Maximizing Efficiency of Data Collection” SNPD Conference 2013

[25] Erik Tjong Kim Sang and Johan Bos “Predicting the 2011 Dutch Senate Election Results with Twitter” Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics , pages 53-60, Avignon, France, April23 - 27 2012. (92012 Association for Computational Linguistics

CURRICULUM VITA

NAME: JONGSUNG YOU

PERMANENT ADDRESS:

Department of Computer and Information Sciences

Towson University

8000 York Road, Towson, Maryland 21252 USA

PROGRAM OF STUDY: Computer Sciences

DEGREE AND DATE TO BE CONFERRED: Bachelor´s degree, 2008

Secondary education:

Shinheung College, gyeonggi-Do, South Korea 2007

Credit Bank University, Seoul, South Korea 2008

Major: Computer Sciences

Professional publications:

Changhyun Byun , Hyeoncheol Lee , Jongsung You , Yanggon Kim “Dynamic Seed Analysis in a Social Network for Maximizing Efficiency of Data Collection” SNPD Conference 2013