Data Mining for Social Media Analysis:Using Twitter to Predict the 2016 US Presidential Election

Kabir Ismail Umar Department of Information Technology, ModibboAdama University of Technology
Yolakabir.ismail@mautech.edu.ng

Fatima Chiroma Software Development Department, American University of Nigeria
Fatima.chiroma@aun.edu.ng

ABSTRACT: Social media is a social platform that is made up of people who are connected by several interdependencies. Social media has changed the nature of information in terms of availability, importance and volume. Through social media like Twitter and Facebook, participants reveal personal information that has real values, as they can be extracted and mined to improve decision making. In this paper, a small sized data was extracted from twitter and analysed using a data mining classification algorithm known as sentiment analysis (also referred to as opinion extraction, opinion mining, sentiment mining and subjective analysis). Sentiment analysis is a technique used by several researchers to measure the emotions of social media participants in online text, in our case it was used to determine the opinion of people or participants towards the 2016 US presidential candidates. Whereas data mining is the technique of discovering and extracting useful information from large data sets or databases. Additionally, the result of the analysis was used to predict the outcome of the aforementioned election.

Key Words – Social Media, Social Media Analysis, Twitter, Sentiment Analysis, Data mining, Election Prediction.

INTRODUCTION

In today's world, one of the most important needs of our daily live is the internet and the social media has massively contributed it. As according to Jamshidi (2008), the web connects millions of people together and provides access to massive resources on the web, making it the largest data repository in the world. Schmidt stated that, "we create as much information in two days now as we did from the dawn of man through 2003" (Siegler, 2010) and most of the data are user-generated from Social media.

According to Grahl (n.d.), there are different types of social media namely Social Networks, Bookmarking sites, Social News, Blog Comments and Forums, Media Sharing and Microblogging. Table 1 show the top five social media platforms (Ebizmba, n.d.):

Name	Type	Total Users
Facebook	Social Networks	1,100,000,000
Twitter	Microblogging	310,000,000
LinkedIn	Social Networks	255,000,000
Pinterest	Social Networks	250,000,000
Google Plus	Social Networks	120,000,000

Table 1: Top Five Social Media Platforms

Since social media is usually formed and constructed by daily and continuous communication between participants, we have decided to investigate its potential in predicting real-world outcome. That is, using the information posted by social media participants to detect their sentiment or opinion about the different 2016 US

presidential candidates. The sentiment will then be used to predict the outcome of the election.

However, this paper will strictly use data from twitter for the analysis. Twitter is a microblogging and social networking platform that allows participants to send short text updates known as "tweets". According to Andrews et al (2016), the following (table 2) are the current US presidential candidates:

Party	Candidate
Democratic	Hillary Clinton
	Bernie Sanders
Republican	Ted Cruz
	John Kasich
	Donald Trump

Table 2: 2016 US Presidential Candidates

The rest of the paper will be structured as follows: In section 2, we will give a background on data mining , social media, data mining techniques and algorithms, and how data mining can be applied to social media to make predictions of the 2016 US presidential election as well as some of the related works that have been done by several researchers; section 3 will describe the methodology we have used i.e. the data collection and data analysis, which will include using some data mining classification algorithms such as sentiment analysis as well as the R package; section 4 will present the experimental results of the analysis carried out in section 3 and our findings from the analysis as well as the prediction results, which will be done using a prediction equation or model; section 5 contains the paper conclusion as well as future works and the last section, section 6 contains the list of the references used for the research/paper.

BACKGROUND

Social media is a platform that allows participants or people to share contents. Sharing contents seem to be a very important part of our daily lives as we are subjective creatures and our opinions are important to us. Different social media platforms have different methods of content sharing. For example, twitter is a social media specifically a social networking site but also a microblogging site whose method of information or content sharing is allowing participants to post short text updates known as tweets. Most of the tweets are opinions or sentiments of participants which has great values as according to Gayo-Avello (2013), twitter data can and "have been mined to determine the public opinion on several topics" including pre-electoral and electoral polls.

However, to mine twitter data, data mining techniques need to be applied. Hand et al defined data mining as a technique used to discover and extract useful information from large databases or data sets (Hand, 2001). Data mining can be applied to different domains such as market analysis, fraud detection, and election prediction. There are five key types of data mining, although some refer to the types as techniques. Chamatkar et al (2014) and Brown (2012) defined them as follows:

- **Association Rule:** is the study of the frequency of items showing up together or it is the correlation between items of the same type for pattern identification.
- **Classification:** is the organization of data in a class or set by identifying its different attributes.
- **Clustering:** this is similar to classification but unlike classification, it is the duty of the clustering algorithm to discover or identify classes that are acceptable.

- **Sequential Patterns:** is the identification of similar events that occur regularly or the identification of behavioural pattern of similar events.
- **Decision Trees:** is also similar to classification, additionally it can be used as a selection criteria of data that are specific within a structure.

According to Totewar (2012), two types of tasks can be performed using data mining and each of the aforementioned techniques or types fall into one of the task category as can be seen in table 3,

Task	Definition	Techniques
Description	To identify or detect human-interpretable patterns that describes the data	Clustering
		Association Rule
		Sequential Patterns
Prediction	To predict the unknowns or future values using some variables.	Classification
		Decision Trees

Table 3: Data Mining Techniques Category

Some of these techniques have been used for social network mining such as classification, association rule and sequential patterns by several researchers like Nandi and Das (Nandi et al, 2013). Furthermore, data mining has several processes that needs to be applied to get an effective result as can be seen in figure 1:

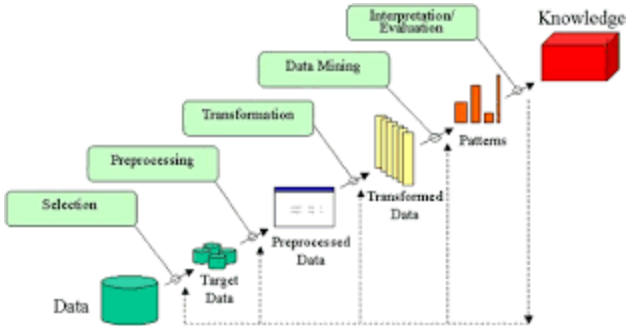

Figure 1: Data Mining Processes (Source: www2.cs.uregina.ca)

To briefly explain the processes in association with this research, we first need to know what exactly we want to mine and for what purpose. In our case, we are looking to mine tweets which contains the opinion of participants towards the US presidential candidates so as to make a prediction on which candidate is most likely to win the general election. After which the required data is collected from twitter and cleaned or pre-processed to get rid of dirty data i.e. noisy, inconsistent and incomplete data, for a more structured, accurate and complete data. Additionally, the data will be transformed from the source (twitter) format to a format that is supported by the tool we will use for the data mining (which is the R package or tool). All of these processes are done before applying data mining techniques such as sentiment analysis, which is a classification algorithm that will be used to classify the transformed data into sentiments. This will enable us to apply a model that will predict the election outcome. In general, the application of the process to our problem is known as data mining application.

Furthermore, Sentiment analysis is an integral part of this research, as it is the point in which a data mining technique is applied. Sentiments are feelings, attitudes, emotions or opinions such as good or bad, positive or negative, like or dislike. While sentiment analysis, also known as opinion mining, is a task that involves extracting

individual feedback information from authentic sources so as to detect their textual opinions (that they have written in digital format)(Choy et al, 2012). These textual opinions are then classified (into good or bad, positive or negative etc.) and the score aggregation method is applied to calculate the score and add the results of matching sentiments. A common use for this technology is to discover how people feel about a particular topic (Lexalytics, n.d.). According to You et al (2015), researchers have largely relied on textual sentiment analysis to develop systems to predict political elections.

"Election isthe formal process of selecting a person for public office or of accepting or rejecting a political proposition by voting" (Gibbins, n.d.). There are a number of previous works that used twitter data to make prediction of elections outcome. Researchers like Choy et al (2012) used twitter data and sentiment analysis to predict the Singapore presidential election 2011. They have claimed that since the successful campaign of the President Obama in 2008, social media platforms such as Facebook and twitter "have catapulted to greet success as the leading platform to engage voters" (Choy, 2012). Research interest in using social media like twitter to make predictions specifically elections, is increasing due to its importance.

Although, Gayo-Avello believes that twitter helps understand the degree of support given to a party but insisted that a twitter user's political leaning does not imply a vote in a given election (2011). He does have a point; however the rapid evolution of technology will lead to the implementation or development of several models that will hopefully eradicate Gayo-Avelo's concerns. Although, researchers like Shi et al, also share Gayo-Avello's concerns as they've stated that using the volume of tweets with or without sentimental analysis is not enough to capture public opinions. They do however believe that twitter has a potential and agreed that more sophisticated algorithms and models needs to be developed to make prediction successful (Shi, 2012).

Additionally, not all researchers agree with Gayo's claim. For example, researchers like Bermingham et al (2011) who had a 65% prediction accuracy for the Irish general election of 2011 and Wang et al (2012) who also had a 59% prediction accuracy for the 2012 US Presidential election, believe that social media has the potential and power to be used as a medium for election prediction. Moreover, Twitter has attracted a lot of corporate attention because of the huge potential it provides for viral marketing (Asur, 2010). As such, participants who share their opinions about elections or electoral candidates on Social Medias like twitter are people who are likely to vote in the general elections or have an influence on those that will vote.

Additionally, Nandi et al (2014) have identified key research issues in social network mining such as influence propagation, behaviour and mood analysis, recommender systems, predicting trust and distrust among individuals, and opinion mining (Nandi, 2014) to name few. Although, one issue that most researchers are genuinely concerned about is the accuracy of the prediction when applied to opinion mining, which is affected by factors like the biasness of tweets, the size of the data and the effectiveness of existing electoral prediction models or equations. You et al (2015), also believes that people not only use textual messages to express their opinion but they also use images and videos, which is a unique challenge for information retrieval and processing as

current models or approaches rely on natural languages or textual opinions.

Conclusively, Nandi et al (2014) strongly believes that algorithms and techniques can still be developed to improve accuracy and speed of social network mining. While Shi et al (2012) on the other hand has emphasised on the need to understand the influence of different lexicons (vocabulary of an individual) by applying machine learning techniques while also been able to integrate our understanding of political dynamic discussions in social media. They believe this will hugely improve the accuracy of election predictions using social media like twitter.

METHODOLOGY

To predict the 2016 US president election, two categorical steps were used for the methodology; the data collection and data analysis.

Data Collection

The data collection step is the initial phase in the research, where data is collected from twitter. Since twitter has some restriction on sharing participants confidential information, they have instead made some provision for developers or researchers to analyse data without viewing its content. This provision requires setting up a twitter developer account with twitter to get authorization codes or access tokens to independently run analysis on tweets.

Additionally, different tools can be used with the access tokens such as RapidMiner, WEKA and R, to name a few. However, the choice for this research is the R package. Using the R package, access tokens and some set of libraries such as the twitteR, plyr, devtools and stringr, data was collected using the presidential candidates' campaign hashtags and a maximum of 10,000 tweets was collected per candidate on the 3rd of April 2016.

The candidates were selected based on the result of the ongoing primary elections. According to Reston (2016), the leading candidates for the democratic and republican parties as at the compilation of this research are Hillary Clinton and Donald Trump respectively.

Furthermore, in addition to collecting the tweets, the word bank which is a collection or database of words has been collected from Michael Herman's GitHub. The word bank is required as it is what will be used to determine whether a participant's sentiment is positive, negative or neutral. Therefore, a total of 2006 positive words and 4378 negative words were collected from the word bank for analysis on the tweets by comparing each word in the tweet extracted to the positive and negative word bank, and then using a sentiment analysis algorithm to determine the overall sentiment or opinion of the twitter users or participants about the different presidential candidates.

During the data collection, we assumed that the participants may take part in the general election or have an influence on those who will vote therefore we did not categorize the participants by age or gender (as we also do not have access to those information). Instead, the data was collected from the first 10,000 tweets for each candidate. As can be seen in table 4:

Candidates	Search Keywords (Hashtags)	Total Tweets
Donald Trump	#trump2016	10,000
Hillary Clinton	#hillary2016	10,000

Table 4: Candidates' Hashtags

Data Analysis

Since the data collected was automatically hidden by twitter for privacy reasons, data pre-processing or cleaning was therefore impossible (as we cannot see the physical data). Instead, the author relied on the capabilities of the R packages and sentiment analysis algorithm to make sense of the collected data.

Therefore, in addition to the sentiment analysis algorithm the R package was also used for the analysis by calling the different functions in the sentiment analysis algorithm. One of the functions is as shown below;

```
analysis=score.sentiment(tweets.text,pos, neg, .progress='no
```

This function was used to compare the tweets with the negative and positive word bank, and cached it in the analysis variable for further or additional analysis such as to get the mean, median, histogram etc.

Furthermore, the word bank (collected) was loaded into the R package. This way the algorithm was able to identify and classify the sentiments of each of the 20,000 tweets into the positive, negative or neutral sentiment per candidates. Table 5 shows the result of the analysis:

Candidate	Total Tweets	Negative	Neutral	Positive	Mean
Hillary Clinton	10,000	2313	5378	2309	-0.02
Donald Trump	10,000	878	1418	7704	1.32

Table 5: Analysis Results

From the result of the analysis, the author then tried to predict the outcome of the election using the polarity lexicon model modified by Gayo-Avello et al. as can be seen in section 4.

EXPERIMENTAL RESULTS

Based on the methodology specified in section 3, the data analysed in table 5 was used to make a prediction by applying the Polarity Lexicon model modified by Gayo-Avello. The model or equation as shown below was used by Gayo-Avello et al to make a prediction of the 2010 US congressional elections (Gayo-Avello, 2011):

Equation 1: Modified Polarity Lexicon

$$(c_1) = \frac{pos(c_1) + neg(c_2)}{pos(c_1) + neg(c_1) + pos(c_2) + neg(c_2)}$$

Table 6gives a description of Equation 1:

Description	
c_1	Candidate 1
c_2	Candidate 2
$pos(c_1)$	Positive words for candidate 1
$pos(c_2)$	Positive words for candidate 2
$neg(c_1)$	Negative words for candidate 1
$neg(c_2)$	Negative words for candidate 2

Table 6: Equation Description

However, the equation does not use the neutral tweets as they don't express a candidate preference. It also does not accommodate more than two candidates; herefore it cannot be used to make a rediction for the primary elections which ost times have more than two candidates.

The prediction was calculated as follows (table 7):

Candidate 1: Hillary Clinton
$c_1 = \frac{2309 + 878}{2309+2313+7704+878} = \frac{3187}{13204} = 0.2414$
Candidate 2: Donald Trump

$c_2 = \frac{7704 + 2313}{7704+878+2309+2313} = \frac{10017}{13204} = 0.7586$
Prediction Result
Hillary Clinton = 0.241 Donald Trump = 0.759

Table 7: Prediction Calculations

From the calculations, it can be evidently seen that Donald Trump has the higher probability with 0.759 compared to Hillary Clinton's with 0.241. Although, the accuracy of the prediction cannot be verified until the general election result is out. We can however proof the correctness of the probability by applying the probability function conditions (Jones 1999) as can be seen in table 8:

Conditions	Proof	Description
All the probabilities must be between 0 and 1 inclusive	0.241 and 0.759	Both probabilities are less than 1 but greater than 0
The sum of the probabilities of the outcomes must be equal to 1	0.241 + 0.759 = 1	Both probabilities sum up to 1

Table 8: Probability Function Conditions

Conclusively, from the calculation in table 7 Donald Trump is the most likely candidate to when the general election.

CONCLUSION

In this research, we were able to show how social media like twitter can be used to make prediction of future outcomes such as elections. Specifically by using R, to extract the sentiment or views of people who are likely to vote in the general election or have an influence on those who will vote; and Sentiment Analysis, to classify their sentiment. Additionally, a Polarity Lexicon model that was modified by Gayo-Avello et al was applied to the result of the analysis to make prediction of the 2016 US presidential elections.

However, as earlier mentions in section 4, the accuracy of the result can only be verified after the election. Therefore, the result of the analysis should be regarded as informative rather than conclusive. Also, as a future work, the model used can be modified to accommodate more than two candidates so that it can be applied to the primary elections not just the general elections or primary elections that have only two candidates. However, the modification of the model will depend on the accuracy of the current 2016 US presidential election prediction.

REFERENCES

(n, d. E. (n.d.). Top 15 most pupolar networking Sites. *Ebiz MBA Guide*. Retrieved April 2, 2016, from http://ebizmba.com/article/social-networking-websites

(n, d. L. (n.d.). Sentiment Analysis. Retrieved April 6, 2016, from

http://www.lexalytics.com/technology/sentiment

Brown, M. (2012, December 11). Data mining Techniques. Retrieved April 5, 2016, from http://www.ibm.com/developerworks/liberary/ba-data-mining-techniques/

Butey, A. C. (2014). Importance of data mining with Different types of Data Applicatons and Challenging Areas. *International Journal of Engineering Research and Application*, 38-41.

Chakrabarti, S. (2003). *Mining the Web: Discovering Knowledge from Hypertext Data.* Caifornia: Morgan Kaufmann Publishers.

College, R. C. (2016). Retrieved from http://people.richland.edu/james/lecture/m170/ch06-prb.html

Das, A. N. (2013). A Survey on Using Data Mining Techniques for Online Social Network Analysis. *International Journal of Computer Science* , 52-57.

Das, G. N. (2014). Online Social Network Mining: Current Trends And Research Issues . *International Journal Research in ENgineering and Technology*, 346-350.

Ferrara, E. (2012). *Mining Analysis of Social Networks, PhD thesis.* Messina.

Gayo-Avello, D. (2013). A Meta analysis of State-of-the-Art electoral Prediction from Twitter Data. 649-679.

Gibbins, R. (2016, Retrieved on April 6). Election. Retrieved from www.britannica.com/topic/election-political-science

Herman, M. (2016, March 19). Wordbanks. Retrieved from http://github.com/mjhea0/twitter-sentiment-analysis/tree/master/wordbanks

Hurbarman, S. A. (n.d.). Predicting the future With Social Media. *WI IAT "10 preceedings of the 2010 IEEE/WIC/ACM international conference on Web Intelligence and Intelligence Agent Technology*, (pp. 492-499). Washington"IEEE.

Jamshdi, B. (2008). *Web Usability in B2B Websites, User' Perspective.* Lulea: University of Technology Press.

Li, C. T. (2011). *User Level Sentiment Analysis Incorporating Social Networks KDD '11*. California: ACM.

Liu, X. H. (2013). Unsupervised Sentiment Analysis with Emotinal Signals. *International World Wide Web Conference* .

Metaxas, D. G. (2011). Limits of Electoral Predictions Using Twitter. *Proceedings of the fifth Internationaql AAAI Conferenceon Weblogs and Social Media.*

Reston, M. (n.d.). 2016 Election. Retrieved April 6, 2016, from http://edition.cnn.com/specials/politics/2016-election

S, H. W., & Narayanan. (2012). *A System for Realtime Twitter sentiment analysis of 2012 US Presidential Election cycle.ACL '12 Proceedings of the ACL 2012 System.* Jeju Island: Association of Computational Linguistics.

Schmidt, M. S. (n.d.). Every 2 Days we create as much Information as we Did Up to 2003. Retrieved from

http://techcrunch.com/2010/08/04/schmidt-data

Shung, M. C. (2012). A sentiment analysis of Singapore Presidential Election 2011 Using Twitter data with census Correction.

Smeaton, A. B. (2011). On Using Twitter to mointor Political sentiment and Predict Election results. *Proceedings of the workshop on sentiment Analysis where AI meets Psychology.*

Smyth, D. H. (2001). *Principles of Data Mining.* Massachusetts: MIT Press.

Spoelstra, L. S. (2012). Prediction US Primary Elections With Twitter.

Totewar, A. (n.d.). Data mining: Concepts and Techniques. Retrieved April 5, 2016, from http://www.slideshare.net/akannshat/data-mining-15329899

You, Q., Luo, J., Jin, H., & Yang, J. (2015). Rubost Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. Association for the Advacement of Artificial Intelligence.

Yourish, W. A. (2016, Retrieved on March 19). Who is running for President? Retrieved from http://nytimes.com/interactive/2016/us/elections/2016-presidential-candidates.html