PREDICTING ELECTION OUTCOME FROM SOCIAL MEDIA DATA

Badhan Chandra Das and Md Musfique Anwar

Department of Computer Science and Engineering, Jahangirnagar University, Bangladesh.
badhan0951@gmail.com
manwar@juniv.edu

ABSTRACT

In this era of technology, enormous Online Social Networking Sites (OSNs) have arisen as a medium of expressing any opinions, thoughts towards anything even support their status against any social or political matter at the same time. Nowadays, people connected to those networks are more likely to prefer to employ themselves utilizing these online platforms to exhibit their standings upon any political organizations participating in the election throughout the whole election period. The aim of this paper is to predict the outcome of the election by engaging the tweets posted on Twitter pertaining to the Australian federal election-2019 held on May 18, 2019. We aggregated two efficacious techniques in order to extract the information from the tweet data to count a virtual vote for each corresponding political group. The original results of the election closely match the findings of our investigation, published by the Australian Electoral Commission.

KEYWORDS

Election Prediction, Social Network Analysis, Twitter, Regular Expression.

1. INTRODUCTION

Over the last few years, OSNs have become a reliable and effective medium of communication and source of important information and news about daily life. OSNs and other micro-blogging sites are such platforms which not only provides the option to be connected with their friends and family members, colleagues, acquaintances, fans and their other well-wishers but it also offers the facility to its users to share different contents like as pictures, texts, videos and others as well. Approximately, 3.09 billion people are connected to various microblogging platforms including Facebook, Twitter, Instagram, Google+, etc [15]. These vast number of people frequently use these online platforms to share their opinion regarding the different affairs of their regular life. In recent years, these OSNs have appeared one of the popular research fields among many prominent researchers over the world [1]. Extracting meaningful information utilizing data of those platforms is a notable task of their research. Twitter is such an online social platform where anyone can write their thoughts about any issue in 140 characters and spread it among their connections. Now it has become one of the greatest origins of news, with above 200 million active users [2].

People share their opinion on various topics, for instance, business, sports, politics, entertainment, literature and culture and so on. During election season concerned people often post tweets on Twitter regarding different political parties including their success, failure, manifestos, positive, negative aspects of their activities or about any specific leader of the corresponding party and also several other stuff. The activity of voters on micro-blogs reflect

their sentiment to any specificpolitical group. Their tweets regarding any political party can be in favour of them or against them. These micro-blogs often play a significant role not only in terms of understanding of public intuition towards any party taking part in the election but also promoting those political groups. Having an aim of winning the election, it is very essential to prepare a manifesto and set the strategy to boost up promotional activities in accordance with the demand of mass people. Oftentimes, the turnout of an election can be predicted approximately from these OSN posts by analyzing the texts of that post on the basis of various relevant parameters. That is why forecasting election outcomes is one of the major and trending tasks of the eminent researchers of text processing. In this paper, we introduce a very simple and straightforward technique of predicting election results based on tweets counts. We have taken into count those tweets, which contain political party names, abbreviations of the party name, any of their leaders' names, the moto of that party and all possible ways by which any corresponding party can be mentioned so that the calculation of estimation of election produces higher accuracy. In brief, the basic approaches of our contributions are as following:

- We propose an easy and novel method of counting system by two-way extraction of the virtual support from twitter mentions of political parties during the election season.
- We perform a comprehensive experiment on a real-world Twitter dataset collected from Kaggle [16] containing 1.8 million tweets collected throughout the election period.
- We explain the techniques followed for recognizing the mention of any specific party in a particular tweet and also present the comparison of the result of our predicted outcomes with the original results of the election.

The remainder of this paper consists of as follows. We describe some relevant works already done in this field in Section 2. In Section 3, we explain the formation of political parties. Section 4 will contain the methodologies used to perform the experiment and the computational equations in order to calculate the election outcome. Section 5 will describe the validation of the experimental results of forecasting as well as make some discussion concerning the outcomes of our presented prophecy. Lastly, the paper puts an end with the conclusion in section 6.

2. RELATED WORKS

Among a bunch of research fields in Text Mining and Natural Language Processing (NLP), forecasting election results, as well as analysis of the public concern on politics, have able to seek the attention of renowned researchers over the world. Earlier works reported some notable works in this field. Besides, political parties are always eager to discover public opinion about them, hence, research on this specific arena has become really essential [3].

Basically, previous investigations have been explored this field from several points of view as follows.

2.1. Analyzing Public Perspective from Social Media

Karami et. al. [4] reported an approach of extracting public opinion in order to explore the discussion of the economic issue during the election period 2012 US presidential election on Twitter. Another Twitter dataset was employed to study public opinion in the case of UK General Election 2010 [5]. Studies have done on predicting the political preferences of users of twitter networks by their online actions [7]. Apart from political issues public opinion mining tasks have been done for several other purposes e.g. decision making for business strategy [17] [18], Kim et. al. have shown how reviews of social media can be a determinant of improvement of restaurants'

performance [19]. Gadekallu et. al. worked on an IMDB data and showed applications of sentiment analysis e.g. determining marketing strategy, improving campaign success, etc [20].

2.2. Election Prediction Using Sentiment Analysis

Sentiment analysis is one of the popular manners of predicting elections among researchers. This approach has been used by the researchers in order to predict elections of several countries. For instance, Indonesia Presidential election [9], The Nigeria Presidential Election 2019 [8], French Election [13]. In accordance with that, Heredia et. al. [3] proposed a prediction technique incorporating location-based sentiment analysis of U.S. Presidential Election 2016. Oyebode et. al. [8] claim a new lexicon-based approach named VADER-EXT which outperforms two other sentiment lexicon library-based approaches known as VADER and Textblob.

2.3. Mixed Approaches

Karami et. al. [4] combined both sentiment analysis and topic modeling approach to find out meaningful information from tweets. Heredia et. al. [6] reported an approach of predicting election of a mixed approach which includes both polling and sentiment analysis. Unankard et. al. [10] introduced a technique of forecasting the election result by associating both sub-event and sentiment analysis. Sanders et. al. [12] in 2016, used demographic information to predict Dutch election results, again, in 2019, asserted that, demographic statistics do not really help in forecasting the elections and proposed a simple tweet count method for doing so [11]. Sanga et. al. proposed a Bayesian network model on Indian general election, 2019 [21].

All the approaches of forecasting the turnout of election cited above are a bit complex and time consuming since they need to employ several lexicon libraries and probabilistic theorem. Our proposed method consists of a very simple and effective procedure for performing the above task.

3. FORMATION OF AUSTRALIAN POLITICAL PARTIES

In this paper, we are about to report an investigation on a dataset of Australian federal election 2019 which took place on 18th May 2019. Several parties participated in the election in order to form the government. A group of parties often make an alliance between them to take part in elections as well as forming the government and for working together later. Making such collaboration between political groups like this is generally known as coalition. The Australian federal election has also seen such a coalition. Table 1 shows the names of the major parties, coalitions, and their leader names.

Table 1. Major Political Parties, their Coalitions and the leaders

Major Political	Parties Aliened With	Leaders Name
Liberal National Coalition	Liberal Party of Australia	Scott Morrison
	Liberal National Party of Queensland	Deb Frecklington
	National Party of Australia	Michael McCormack
	Country Liberal Party	Gary Higgins
Australian Labor Party	No Coalition	Anthony Albanese
The Australian Greens	No Coalition	Richard Di Natale

4. PROPOSED METHODOLOGY

4.1. Counting Procedure of Tweets

A huge amount of Twitter dataset which contains around 1.8 Million tweets collected from 10 May 2019 to 20 May 2019. As the election date comes near, the frequency of posting tweets regarding election increases rapidly. A political party in a tweet can be mentioned in various ways. Sometimes a party name may be a bit large, besides, writing the full name with the proper spelling can be daunting as well as time-consuming. This is why referring the full name of any political party in a tweet is not convenient always. Moreover, it is not mandatory to mention the full official name of any organization on such a casual online platform like Twitter. Similarly, from our observation of the dataset, we noticed that most of the users who posted about any political party on Twitter did not write the complete formal name of the party always. People commonly used acronyms, a portion of the name by which a party can be recognized, the name of the leader of the corresponding party or any philosophy, ideology or doctrine of that respective political group to espouse them. So, determining the tweets from a huge amount of data contains names of the specific political wing, needs a generalized method that can extract only those data pertaining to analogous information regarding that party.

Considering all those points above, we have chosen the following approaches to construct the counting process more efficient.

4.1.1. Regular Expressions for Party Mentions

Regular Expression is such a technique to find any specific or any conditional pattern, characters, numbers or symbols in a text. Utilizing the above technique to identify the party names in the tweets by taking almost all possible cases into account through which a party can be cited in an instance of a tweet. For example, the regular expression for the first instance of Table 2 will detect the mention of the party Liberal National Coalition as Liberal National Coalition, LNC, L.N.C., Liberal Party and all other possible cases. The Regular Expressions we employed in order to find party mentions shown in the following table.

Table 2. Regular Expressions to find party mentions in tweets

Name of the Major Political Parties	Regular Expressions used for possible cases of mentions of political groups in tweets
Liberal National Coalition	(([[^L|l][.|_]?]*((IBERAL|iberal))?*)?([[N|n][.|_]?]*((ATIONAL|ational))? *)(([[C|c][.|-|_]?]*((oalition|OALITION))?))? *(([[^L|l][.|-|_]?]*((IBERAL|iberal))? *)?)?([[^P|p][.|-|_]?]*((ARTY|arty))? *)? *(([[C|c][.|-|_]?]*((oalition|OALITION))?))? *)
Australian Labor Party	(([[^A|a][.|-|_]?]*((ustralian*|USTRALIAN)*))?\s*([[L|l][.|-|_]?]*((abor|ABOR))? *)([[P|p][.|-|_]?]*((arty|ARTY))?) *([[^A|a][.|-|_]?]*((ustralian|USTRALIAN)))?)
The Australian Greens	(([[T|t][.|-|_]?]*((HE|he))? *)?([[A|a][.|-|_]?]*(USTRALIAN|ustralian))? *([[G|g][.|-|_]?]*((REENS|reens))? *)([[P|p][.|-|_]?]*((ARTY|arty))? *)?)

4.1.2. Search for Specific Relevant Keywords

Sometimes it is seen that mass public quote the popular morals, taglines, ideologies or the name of the leader or the founder of any political group to show their enthusiasm towards the party. As a result, any generalized strategy is unable to recognize these types of implicit mentions in the tweets. Taking this issue into account, in order to resolve this problem, we manually searched according to the above-mentioned keys in the tweets. We engaged str.contains(‘’) function offered by Pandas library in python version 3.7.

After having outputs from both operations, we aggregate both of the outputs to perform the computational experiments to get the prediction result.

4.2. Parameters Consideration

i. We have ignored the location, occupation and other demographic features of twitter users since this paper introduces a very simple approach to forecast election results.
ii. Removal of redundant instances and elimination of stopwords has been done at the early stage of the experiment.
iii. Retweets and comments were not taken into counts because retweeting not denote the support for the original tweet post every time. Along with this, URLs in a tweet don't hold much information in terms of advocation for any particular group. For this reason, URLs also get outweighed our attention to be counted as a parameter.
iv. As we stated earlier, the tendency of tweeting regarding election rises considerably when the date of election comes closer. Figure 1 demonstrates this statement graphically, where X-axis denotes the date of the month of May 2019 and the Y-axis indicates the number of tweets.
v. Our consideration was to measure the mentions for the top three parties received 96.7% votes in the actual election held on 18th May 2019 in Australia.

Figure 1: Graphical representation of party mentions during election period

4.3. Average Tweets Count and Percentage

Let *MC(X)(i)* be the counts for mentions on behalf of party *X* on the day *i* and N is the sum of the day counts. U_n is the number of unique users tweets in favour of party *X*, then, the average tweet counts per user mention among all the tweets for party *X* is,

$$ATCP(X) = \frac{\sum_{i=1}^{N} MC(X)(i)}{U_n} \qquad 1$$

Where *ATCP(X)* denotes the average tweets counts per user mention in the tweet for party *X*,

$$PREC(X) = \frac{Ceil(ATCP(X))}{\sum_{j=1}^{P} Ceil(ATCP(j))} \qquad 2$$

If the number of political parties is *P*, then the party *X* acquires the percentage of mentions denoted by *PREC(X)* shown in Eq. 2. Where the numerator is computed after applying ceiling function on the values yielding from Eq. 1 and denominator represents the summation of all average tweet counts for *P* parties. Ceiling function of *x* is the value gives the smallest integer which is greater than or equal *x*. For example, *Ceil(3.5)* will provide *4* as output.

5. EXPERIMENTAL RESULT AND DISCUSSION

The evaluation of our proposed framework has been conducted by two of the most common and well recognized metrics named Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which are frequently used to measure accuracy for continuous variables.

Mean Absolute Error (AE) it is a term which determines the measurement of flaws in the assessment compared with original values. It is also known as Absolute Accuracy Error (AAE). MAE is the average of all AEs. In our proposed context, $PREC_{el}(i)$ and $PREC_{mc}(i)$ refer to the percentage of vote earned by party *i* in the election and the percentage of the mention counts for party *i* among the tweets through our suggested system.

$$MAE = \frac{1}{P}\sum_{i=1}^{P}(|PREC_{el}(i) - PREC_{mc}(i)|) \qquad 3$$

Root Mean Squared Error (RMSE) is known as a measurement of standard deviation which denotes how much spread out the predicted value is from the real value. Usually, this procedure is appropriate for determining the standard deviation of the residuals. Residuals are referred to the prediction errors which means how distant the regression line is from the actual data points. Eq. 4 shows how RMSE is calculated.

$$RMSE = \sqrt{\left(\frac{1}{P}\sum_{i=1}^{P}(PREC_{el}(i) - PREC_{mc}(i))^2\right)} \qquad 4$$

The implicit indication behind the above discussion is that the less the value of these two metrics the greater the accuracy of the framework will be.

The comparison between the percentages of our predicted outcome and the real election result represented graphically in Figure 2. We can see clearly that, the flaw of our investigation is very

little for the top two parties according to the real turnout of the election that took place on 18th of May [14].

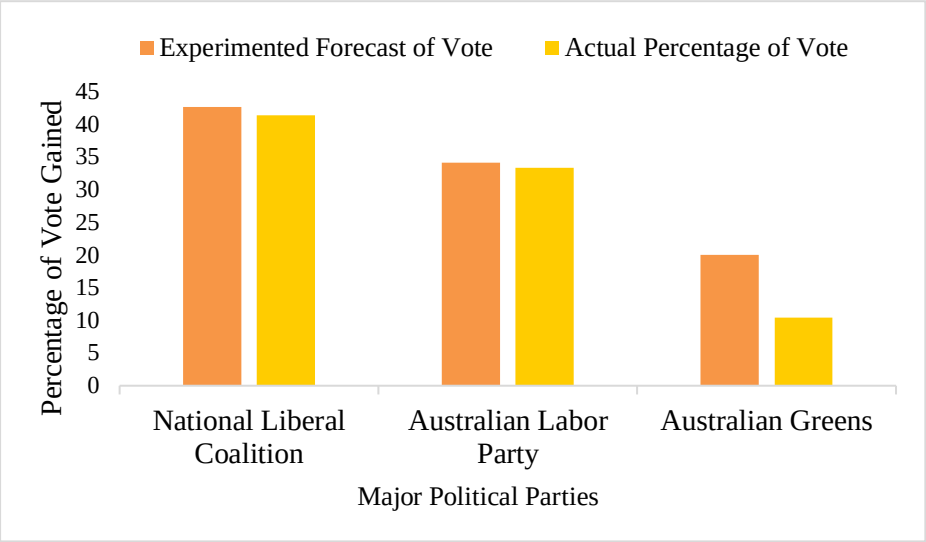

Figure 2: The comparison between the percentage of forecast from tweets and votes of actual election for three major political parties

The correctness of any system grows higher as the value of MAE and RMSE drops and the value of RMSE will always be greater than the value of MAE. Our experimental result follows these ground rules of the two above mentioned accuracy metrics. The value of the MAE and RMSE of the experimental result is shown in Table 3.

Table 3. The Result of Two Evaluation Metrics with respect to the main election

Evaluation Metrics	Australian Federal Election 2019
MAE	3.91%
RMSE	5.63%

6. CONCLUSION

In this paper, we meticulously examined a dataset of Australian federal election 2019 and made a forecast of the percentage of votes gained by each party. The counting procedure followed two useful manners to extract party mentions in an instance of a tweet. The obtained outcome yields our proposed approach closely matches the actual percentage count of the election and gives a little error. The evaluation of the prediction outcomes has executed by two most well-known accuracy metrics.

REFERENCES

[1] Franch, Fabio: (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. In: Journal of Information Technology & Politics, v.10, n.1, pp. 57-71, 2013.

[2] Wang, Lei, and Gan, John Q: Prediction of the 2017 French election based on Twitter data analysis. In: 2017 9th Computer Science and Electronic Engineering (CEEC), pp. 89-93, 2017.

[3] Heredia, Brian and Prusa, Joseph D and Khoshgoftaar, Taghi M: Location-based twitter sentiment analysis for predicting the U.S. 2016 presidential election. In Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31), 2018.

[4] Karami, Amir and Bennett, London S and He, Xiaoyun: Mining public opinion about economic issues: Twitter and the U.S. presidential election. In: International Journal of Strategic Decision Sciences (IJSDS), v.9, n.1, pp. 18-28, 2018.

[5] Anstead, Nick and O'Loughlin, Ben: Social media analysis and public opinion: The 2010 UK general election. In: Journal of Computer-Mediated Communication, v.20, n.2 pp. 204-220, 2015.

[6] Heredia, Brian and Prusa, Joseph D and Khoshgoftaar, Taghi M: Social media for polling and predicting United States election outcome. In: Social Network Analysis and Mining, v.8, n.1, pp. 48-64, 2018

[7] Makazhanov, Aibek and Rafiei, Davood and Waqar, Muhammad: Predicting political preference of Twitter users, In: Social Network Analysis and Mining, v.4, n.1, pp. 193, 2014.

[8] Oyebode, Oladapo and Orji, Rita: Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019, In: 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 140-146, 2019

[9] Budiharto, Widodo and Meiliana, Meiliana: Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis, In: Journal of Big data, v.5, n.1, pp. 51, 2018

[10] Unankard, Sayan and Li, Xue and Sharaf, Mohamed and Zhong, Jiang and Li, Xueming: Predicting elections from social networks based on sub-event detection and sentiment analysis, In: International Conference on Web Information Systems Engineering, pp. 1-16, 2014

[11] Sanders, Eric and van den Bosch, Antal: A Longitudinal Study on Twitter-Based Forecasting of Five Dutch National Elections, In: International Conference on Social Informatics, pp. 128-142, 2019

[12] Sanders, Eric and de Gier, Michelle and van den Bosch, Antal: Using demographics in predicting election results with Twitter, In: International Conference on Social Informatics, pp. 259-268, 2016.

[13] Wang, Lei and Gan, John Q: Prediction of the 2017 French election based on Twitter data analysis, In: 9th Computer Science and Electronic Engineering (CEEC), pp. 89-93, 2017.

[14] AustralianFederal Election Result Party Totals 2019: https://www.abc.net.au/news/elections/federal/2019/results/party-totals [Last Accessed 28 Jan. 2020]

[15] Number of social network users: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [Last Accessed 28 Jan. 2020]

[16] Australian Election 2019 Tweets: https://www.kaggle.com/taniaj/australian-election-2019-tweets [Last Accessed 28 Jan. 2020]

[17] Shen, Chien-Wen and Ho, Jung-Tsung: Public Opinion Toward Social Business from a Social Media Perspective, In: International Conference on Data Mining and Big Data, pp. 555-562, 2018.

[18] He, Wu and Wang, Feng-Kwei and Akula, Vasudeva: Managing extracted knowledge from big social media data for business decision making, In: Journal of Knowledge Management, 2017.

[19] Kim, Woo Gon and Li, Jun Justin and Brymer, Robert A: The impact of social media reviews on restaurant performance: The moderating role of excellence certificate, In: International Journal of Hospitality Management, v.55, pp. 41-51, 2016.

[20] Gadekallu, ThippaReddy and Soni, Akshat and Sarkar, Deeptanu and Kuruva, Lakshmanna: Application of Sentiment Analysis in Movie reviews, In: Sentiment Analysis and Knowledge Discovery in Contemporary Business, pp. 77-90, 2019.

[21] Sanga, Aniruddh and Samuel, Ashirwad and Rathaur, Nidhi and Abimbola, Pelumi and Babbar, Sakshi: Bayesian Prediction on PM Modi's Future in 2019, In: Proceedings of International Conference on Recent Innovations in Computing, pp. 885-897, 2020.

AUTHORS

Badhan Chandra Das is an M.Sc. student of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. He has completed his B.Sc. also in Computer Science and Engineering, from Jahangirnagar University in 2017. His research is concentrated on Data Mining, Social Network Analysis, Natural Language Processing and developing Machine Learning models for Intelligent tools. He has several publications on international conferences. Currently, he is working on a funded research project of University Grants Commission (UGC) of Bangladesh.

Md Musfique Anwar has awarded PhD degree from the Department of Computer Science and Software Engineering, Faculty of Science, Engineering and Technology of Swinburne University of Technology, Melbourne, Australia in 2018. He has received M.Sc. degree from the Department of Intelligence Science and Technology, Graduate School of Informatics of Kyoto University, Japan in 2013 and B.Sc. degree in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka, Bangladesh in 2006. Since 2008, he is a faculty member having current designation Associate Professor in the Department of ComputerScience and Engineering of Jahangirnagar University, Savar, Dhaka, Bangladesh. Currently, his research focuses on Data Mining, Social Network Analysis, Natural Language Processing and Software Engineering. He achieved Best Student Paper award in 29th Australasian Database Conference (ADC) in 2018, Best Poster award in 26th Australasian Database Conference (ADC) in 2015 and Best Poster Paper award in International Workshop on Computer Vision and Intelligent Systems-2019 (IWCVIS2019) in 2019.