Twitter Based Election Prediction and Analysis

Pritee Salunkhe, Sachin Deshmukh

[1,2]*Department of Computer Science and Information Technology,*
Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India

---

Abstract - *Election is conducted to view the public opinion, where group of people choose the candidate by using votes, many methods are used to predict result. many agencies and media companies conduct pre poll survey and expert views to predict result of election. We use twitter data to predict outcome of election by collecting twitter data and analyze it to predict the outcome of the election by analyzing sentiment of twitter data about the candidates. We used lexicon based approach with machine learning to find emotions in twits and predict sentiment score.*

Key Words: Poll, election prediction, microblog, political sentiment

1. INTRODUCTION

An election is a most important part in the democracy. it's the most instrument of democracy wherever the voters communicate with the representatives. Due to their important role in politics, there always has been a big interest in predicting an election outcome. It is the main instrument of democracy where the citizens communicate with the representatives. One vital component in an election is that the election polls/survey.

An opinion poll has existed since the early 19th century, based on [1]. And currently, there are many scientifically proven statistical models to forecast an election, as shown in [2]. But sometimes, even in the developed countries, the polls failed to accurately predict the election outcomes. [3] listed several failed polls result such as in the 1992 British General Elections, the 1998 Quebec Elections, the 2002 and 2007 French presidential elections, the 2004 European elections in Portugal, the 2006 Italian General Elections, and the 2008 Primary Elections in the States.

Lately, it is observed that traditional polls may fail to make an accurate prediction. The scientific community has turned its interest in analyzing web data, such as blog posts or social networks' users' activity as an alternative way to predict election outcomes, hopefully more accurate. Furthermore, traditional polls are too costly, while online information is easy to obtain and freely available. This is an interesting research area that combines politics and social media which both concern today's society. It is interesting to employ technology to solve modern-day challenges.
Trying to resolve the accuracy and high cost problem, we study the possibility of using data from social media as the data source to predict the outcome of an election. Social media has become the most popular communication tool on the internet. Hundreds of millions of messages are being posted every day in the popular social media sites such as Twitter4 and Facebook5. [4] Stated in their paper that social media websites become valuable sources for opinion mining because people post everything, from the details of their daily life, such as the products and services they use, to opinions about current issues such as their political and religious views. The social media providers enable the users to express their feelings or opinions as much as possible to increase the interaction between the users and their sites. This means that the trend on the internet is shifting from the quality and lengthy blog posts to much more numerous short posts that are posted by a lot of people. This trait is very valuable as now we can collect different kind of people's opinions or sentiments from the social web.

One of the social media that allows researchers to use their data is Twitter. Twitter is a microblogging web service that was launched in 2006. Now, it has more than 200 million visitors on a monthly basis and 500 million messages daily. The user of twitter can post a message (tweet) up to 140 characters. The message is then displayed at his/her personal page (timeline). Originally, tweets were intended to post status updates of the user, but these days, tweets can be about every imaginable topic. Based on the research in [5], rather than posting about the user's current status, conversation and endorsement of content are more popular. The advantages of using tweets as a data source are as follows; first, the number of tweets is very huge and they are available to the public. Second, tweets contain the opinion of people including their political view.

2. RELATED WORK

In this section, we are going to discuss related works about predicting the result of an election using Twitter. We noticed that researchers use a different approach regarding this problem. There are researchers who try to discover the political or ideology preference of a user, then relate it to the election and there are others who use selected tweet related to the upcoming election and figure out vote preference of the user using that data.

Different strategies such as profile details, user behavior, Twitter specific feature (reply/re-tweet), user graph and sentiment from tweet content can be used for inferring political leaning. For example, In [6], the authors used tweet containing parties' name in several political events to assign a political/ideological leaning of the user who posted the tweets. Similar to the previous method, [7] used the tweets and retweets of a user regarding a political party to infer the political leaning. [8] Assigned a score to every congress member which a Twitter user is following, then a political preference is assigned based on that score. In [9], the authors compared several features such as user's bio and avatar, posting behavior, linguistic content, follower, reply and retweet. They found out that the combination between user profile and linguistic outperform other feature. They then applied to classify the ethnicity of the user and whether the user is a Starbucks fan, but their result showed that information from user bio is more accurate for classifying Starbucks fan, and user's avatar for classifying user's ethnic.

The second approach is by using selected data just days or weeks prior to the election. The prediction could be derived by comparing the number of tweets mentioning each candidate or by comparing the number of tweets that has positive sentiments towards each candidate. The earliest research stated that the number of tweets mentioning a party reflects the election result was shown in [10] where they found out that the prediction result from Twitter were only better than other. While [11] is the first research in which argued that sentiment detection approach from Twitter can replace the expensive and time intensive polling?

Researchers have tried to compare these two methods, for example, [12] that tried to predict congress and senate election in several states of the US. They showed that though the method is the same, the prediction error can vary greatly. The research also showed that lexicon based sentiment analysis improves the prediction result, but the improvements also vary in different states. Same result was shown in [13] where they predict the result of Irish general election using both methods and [14] which predicts the Italian primary election. All of the research showed that sentiment detection does reduce the error of the prediction result. Because of that, several researchers focused on improving the sentiment analysis, such as[14] and [15] who used more sophisticated sentiment analysis than lexicon based in the US presidential election, France legislative election, and Italy primary election.

Other than using sentiment analysis, the prediction result from Twitter can be improved by using user normalization. This is based on the fact that in an election, one person only has one vote. [16] Implemented this method and showed that the prediction result of 2011 Dutch senate election was improved. [17] Takes further step by adding census correction on the user normalization. [18] Also implemented this method in several South American countries. He collected more than 400 million of tweets, and got a very good result (low difference with the election result) predicting Venezuela presidential election. But when applying in Ecuador and Paraguay presidential election that has much less dataset, the error of the prediction increases significantly.

Other methods proposed by researchers are by:

(1) Utilizing interaction information between potential voter and the candidates and
(2) Creating trend line from the changes in follower of the candidates.

[19] Used interaction information such as the number of interaction, the frequency of interaction, the number of positive and negative terms in the interactions in the Canadian legislative election. The candidates were grouped into four parties, and based on their result; they argued that that the generated content and the behaviour of users during the campaign contain useful knowledge that can be used for predicting the user's preference. [20] Tried to utilize the size of candidates' network (follower in Twitter and friend in Facebook), but the result showed that it was not a good predictor of election results. One interesting result from their research is that despite the huge size of social media, it has small effect on the election results. Therefore, it only makes a difference in a closely contested election.

However, there are several researchers arguing that research in this area is still premature and requires a lot of development before it can give satisfying prediction result. [21] Argued that prediction model using Twitter only able to predict the result from the top candidates/parties and slight variable changes in the model did impact the prediction result. In [22], the authors listed several drawback of the research in this topic such as, most predictions are actually a post-hoc analysis, no commonly accepted way exists for "counting votes", the sentiment analysis methods are not reliable, no data cleansing step, demography and self-selection bias has not been addressed. In [23], in addition to previously stated drawbacks, gave several suggestions such as the importance of geographical and demographical bias, the noise in the social media, the reproducibility of proposed methods, and MAE should be use rather than only winner prediction.

3. METHODOLOGY

3.1 Data Collection

The data collection step is the initial phase in the research, where data is collected from twitter. There are two methods on how to connect and collect tweets from Twitter. The first method is by searching tweets matching to the keywords. The second method is by collecting all the tweets provided by Twitter through streaming API, or all the tweets in a specific language, or all the tweets in a specific location then put all of them into the database.

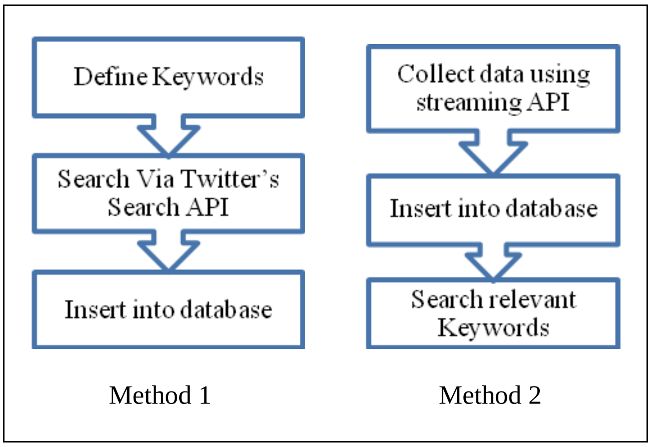

Fig-1: Data Collection Methods

Both methods have their own advantages and disadvantages. For example, the first method requires only small storage as the data are relatively small. The downside is that researcher cannot get data from other keywords (if he needs to) from an earlier time. Twitter allows the search API only for 7 days backwards. This data collection method is suitable if the focus of the research is on the feature extraction or the prediction method. With the second method, researcher can apply many set of keywords to get the best result.

3.2 Preprocessing

Many current methods for text sentiment analysis contain various preprocessing steps of text. One of the most important goals of preprocessing is to enhance the quality of the data by removing noise. Another point is the reduction of the feature space size.

a) Lower Case Conversion:

Because of the many ways people can write the same things down, character data can be difficult to process. String matching is another important criterion of feature selection. For accurate string matching we are converting our complete text into lower case.

b) Removing Punctuations and Removing Numbers:

All punctuations, numbers are also need to remove from reviews to make data clean and neat. Unnecessary commas, question marks, other special symbols get removed in this case. Here, we are not removing dot (.) symbol from our reviews because are splitting our text into sentences.

c) Stemming:

Stemming is that the method of conflating the variant styles of a word into a standard illustration, the stem. For example, the words: "presentation", "presented", "presenting" could all be reduced to a common representation "present". This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented. Stemming in our case helpful in correct words matching and counting case.

d) Striping White Spaces:

In this preprocessing step all text data is cleansed off. All unnecessary white spaces, tabs, newline character get removed from the text.

3.3 Sentiment Analysis:

a) Machine Learning Approach:

There are two approaches of machine learning, supervised and unsupervised. in our research we used supervised machine learning approach.

In supervised machine learning approach there is finite set of classes for classification. Training dataset is also available. Most research papers do not use the neutral class, which makes the classification problem considerably easier, but it is possible to use the neutral class. Given the training data, the system classifies the document by using one of the common classification algorithms such as Support Vector Machine, Naïve Bayes etc. We used naive bays for classification of tweets. We classified tweets into polarity and emotion also using naive bays classifier.

Naive Bayes is a machine learning algorithm for classification problems. It is based on Bayes' probability theorem. It is primarily used for text classification that involves high dimensional knowledge sets.A few examples are spam filtration, sentimental analysis, and classifying news articles.

It is not only known for its simplicity, but also for its effectiveness. It is fast to build models and make predictions with Naive Bayes algorithm.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1)$$

Where,

P(A|B):Probability (conditional probability) of occurrence of event given the event B is true.

P(A) and P(B): Probabilities of the occurrence of event A and B respectively.

P(B|A): Probability of the occurrence of event B given the event A is true.

b) Lexicon Based Approach:

There three main approaches to compile sentiment words. Three main approaches are: manual approach, dictionary-based approach, and corpus-based approach.in our research we used dictionary based approach. We used eleven different variables for classification, that variables are sadness, tentativeness, anxiety, work, anger, certainty, achievement, positive words, negative words, positive hashtag and negative hashtag. We collected various word related to that eleven variable and classified them.

4. RESULTS

We collect data through twitter API, after that we performed preprocessing on that data. For collected data for US and Gujarat Rajya Sabha election we classified polarity. Classifying tweets in three categories positive, negative and neutral.

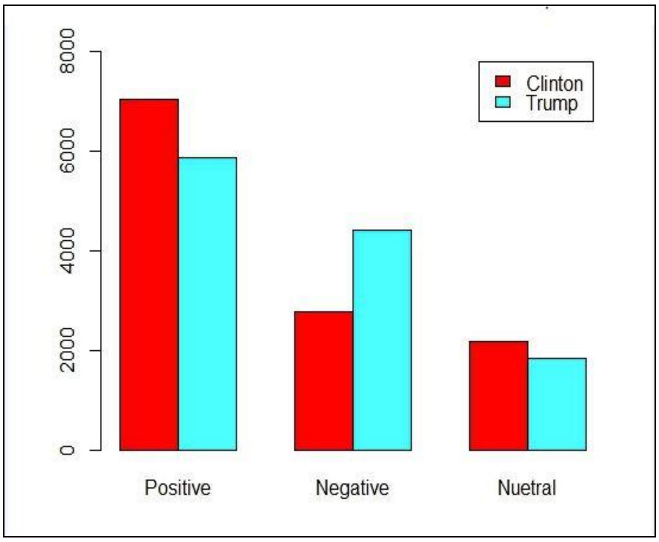

Chart-1: US Election Sentiment Analysis

We observed that in sentiment analysis positive tweets were more in Clinton data than that of trump and vice versa. Neutral tweets are also more for Clinton. As prediction we can predict that Hilary Clinton will win elections from tweeter data we collected. we can see the comparison of positive, negative and neutral tweets for Clinton ad trump in chart-1.

After we take six different variables for emotion analysis. six different variable such as joy, surprise, anger, disgust, fear and sadness.

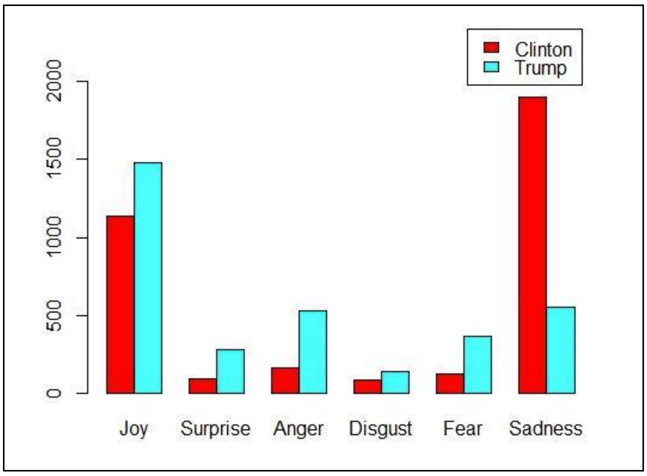

Chart-2: US Election Emotion Analysis

We observed that Sadness and joy were among most expressed emotions in our data. Hillary Clinton's had more tweets about sadness and trump has more expressing joy. Using naïve byes classifier to train our data, we observed that the output observations on tweeter data it shown in chart-2. This suggests that in election trump will get benefit of joyous tweets and trump tweets has more tweets on all other emotions. It can be observed that large part of Clinton tweets express sadness.

We use dictionary based approach, for prediction we used eleven variables. And classify tweets in eleven different variables, that variables are sadness, certainty, tentativeness, anxiety, work, anger, achievement, positive word, negative words, positive hashtag and negative hashtag as in chart-3.

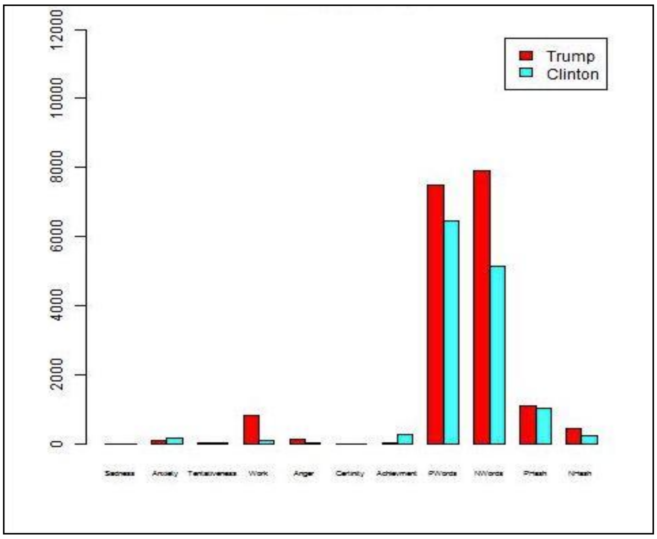

Chart-3: US Election lexicon Analysis

Then we analysis Gujarat election data. We analysis this data for Gujarat Rajyasabha election. There are 4 candidates are part in Gujarat election. Out of four, three candidates are elected and one is not elected. And we classify this data into three categories.

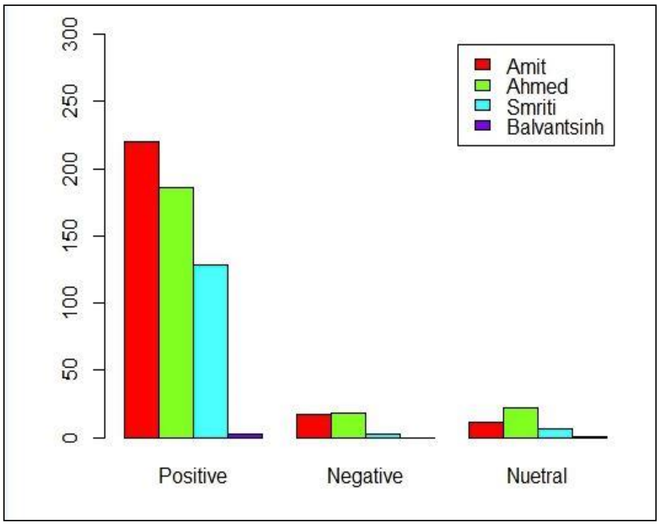

Chart-4: Gujarat Election Sentiment Analysis

In analysis of sentiment analysis in Gujarat elections, we observed that more positive tweets were posted about Amit followed by Ahmed and Smriti. Balvant Sinh had few tweets which were considered positive in nature. Ahmad had more neutral tweets ad negative it can be predicted that Amit is strongest candidate from sentiment analysis data.

We also use six variables for Gujrat election. For Gujarat Rajya sabha election, it was observed that all the Candidates had more tweets expressing joy than any other emotion. Balvant sing had very less tweets to be considered as candidate and predicted not to win the election. Candidate Ahmed had more number of tweets expressing joy and sadness than any other candidate followed by Amit and Smriti respectively. For prediction of result of election we predicted that for 3 seats and 4 candidates Amit, Ahmed and Smriti to win and Balvant Sinh to lose the election.

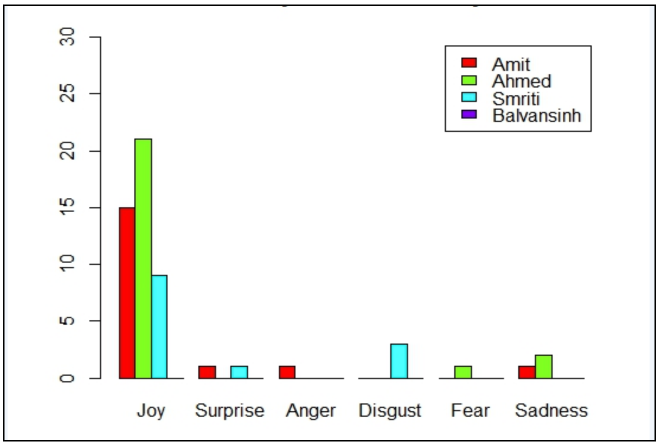

Chart-5: Gujarat Election Emotion Analysis

Finally, we classified Gujarat election tweets into eleven variables. Using that we classified tweets into eleven variables. For that classification we used dictionary based approach, that shown in chart-6.

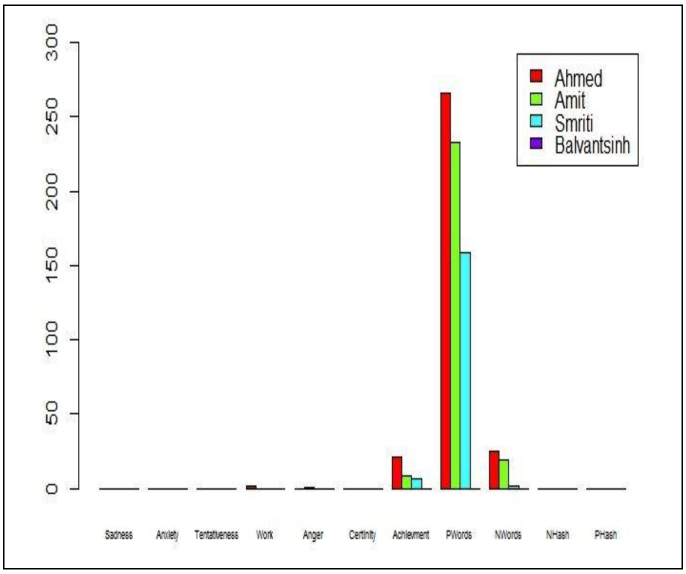

Chart-6: Gujarat Election Lexicon Analysis

5. CONCLUSIONS

In this research, we were able to show how the social media like twitter can be used to make prediction of future outcome such as election 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.

REFERENCES

[1] Hillygus, D. S. (2011). The evolution of election polling in the United States. Public opinion quarterly, 75(5),, 962-981.

[2] Lewis Beck, M. S. (2005). Election forecasting: principles and practice. The British Journal of Politics & International Relations, 7(2), 145-164.

[3] Fumagalli, L. &. (2011). The total survey error paradigm and pre-election polls: The case of the 2006 Italian general elections. ISER Working Paper Series. 2011-29.

[4] Pak, A. &. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC.

[5] Dann, S. (2010). Twitter content classification. First Monday, 15(12).

[6] Wong, F. M. (2013). Quantifying Political Leaning from Tweets and Retweets. ICWSM.

[7] Boutet, A. K. (2012). What's in your Tweets? I know who you supported in the UK 2010 general election. Proceedings of the International AAAI Conference on Weblogs and Social Media.

[8] Golbeck, J. &. (2011). Computing political preference among twitter followers. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[9] Pennacchiotti, M. &. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining , 430-438.

[10] Tumasjan, A. S. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.

[11] O'Connor, B. B. (2010). From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11, 122-129

[12] Gayo-Avello, D. M. (2011). Limits of electoral predictions using twitter. ICWSM.

[13] Bermingham, A. &. (2011). On using Twitter to monitor political sentiment and predict election results.

[14] Ceron, A. C. (2014). Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy. Social Science Computer Review.

[15] Ceron, A. C. (2013). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society, 16(2), 340-358.

[16] Sang, E. T. (2012). Predicting the 2011 dutch senate election results with twitter. the Workshop on Semantic Analysis in Social Media (pp. 53-60). Association for Computational Linguistics.

[17] Choy, M. C. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model. arXiv preprint arXiv:1211.0938.

[18] Gaurav, M. S. (2013). Leveraging candidate popularity on Twitter to predict election outcome.Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM., 7.

[19] Makazhanov, A. R. (2014). Predicting political preference of Twitter users. Social Network Analysis and Mining, 1-15.

[20] Cameron, M. P. (2013). Can Social Media Predict Election Results? Evidence from New Zealand. No. 13/08.

[21] Jungherr, A. J. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe,"predicting elections with twitter: What 140 characters reveal about political sentiment". Social Science Computer Review, 30(2),, 229-234.

[22] Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. Internet Computing, IEEE, 16(6), 91-94.

[23] Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. Social Science Computer Review.