Predictive Analysis on Electoral Poll using micro-blogging (twitter)

Tabraiz Anwer
MS Computing
Shaheed Zulfiar Ali Bhutto Institute of Science and Technology
90 and 100 Clifton
Karachi -75600
tabraizanwer@yahoo.com

Adeel Ahmed
MS Computing
Shaheed Zulfiar Ali Bhutto Institute of Science and Technology
90 and 100 Clifton
Karachi -75600
adeel.ahmed@szabist.edu.pk

***Abstract*— The advent of blogging more and more content is published on the web. Micro blogging is used to convey ideas using short text messages; video links or images .micro blogging platforms allow the public to convey their ideas through new technologies and to convey ideas and news in precise manner. This paper aims to use to text mining algorithm, which is used for gathering relevant information from passages of text, to derive a context through which predictive analysis can be made. The goal will be to use this technique to provide a predictive analysis on the electoral polls using tweets from the micro blogging platform twitter.**

Keywords Terms— Text Mining, Domain Specific Terms Extraction.

I. INTRODUCTION

Social Media is the platform that allow users to express its view worldwide. Social Media has demonstrate its growth and profound influence. Since sharing of opinion and experience via social media, it is an aggregation of different viewpoints and it is related to subject to change with time. Prediction through social Media, if extracted and analyzed properly, it could lead us to helpful prediction of human related issue .This research topic inquires the role of micro-blogging site (twitter) in electoral event. Although candidates and political parties agree on the importance of social media and interested in promoting their presence in social media. Research by Gibson & McAllister, Gueorguieva, Gulati and Williams have examined the importance of social networks and tools of the web 2.0 in the most recent electoral campaigns held in developed democracies.This study seek to contribute to the accumulated knowledge about electoral campaigns and social networks. In election 2008, barack obama change the trend of US election campaign by interacting through social media. His site www.mybarackobama.com, helped him to set records in donations and mobilization.

A. Text Mining

Text Mining refers to the analysis and extraction of information from the text. It involves the process of structuring text (parsing, tokenization ,stemming, generating N-gram), retrieving patterns , finally evaluation and interpretation of output. Our focus is on classification and prediction. These are the widely studied and applied methods in text mining but different problems can also be solved using text-mining techniques.

B. Applications of text mining

Majorly text mining central task is extract information from unstructured text. This field also covers inter-related problems such as manipulation, storage and retrieval of language data.

C. Document Classification

Categorization of unlabeled articles. At the beginning some sample data will be provided with some classes such as “Sports”, “Technology” and “gadgets” to a categorization system and then it will automate itself according to the previous categorized articles submitted earlier at the stage.

D. Information Retrieval

It is mostly related to the online document. We give some attribute to online document that we want to find out that values from the document and document matching attributes are presented as answer of our attributes. Basic concept is to find out similarity between the documents.

E. Clustering Document

It is equivalent to the assigning label needed for the text categorization. It is not much powerful as text classification. But labeling document and making cluster of documents can be insightful for companies wanted to know about which domain have more problems then other.

F. Information Extraction

It is related to extracting structured information from the unstructured text. In most cases it deals with human language processing through Natural language processing.

G. Prediction and Evaluation

This area is generally related to prediction,program learn generalized rules from the example document, that will give correct answer on behalf of generalized rules.

II. MICRO-BLOGGING SITE (TWITTER)

Twitter is most popular micro-blogging site. It is considered as Direct Social Network. User has the list of followers and user can update their status knows as tweets that is consists of 140 characters. This tweets consists of personal messages , information, or links contains of images, links and videos. When user update his status it is shown in his/her profile page also viewed to its subscribers followers.

If message that is posted by user, is forwarded by the other user, this term is known as Retweet and it is popular mean propagating interested messages. Twitter has attracted lot of organizations towards itself for marketing its products as it has potential for viral marketing. Due to his vast usage it has been used by news agencies, to filter its news updates. Also numbers of organization is using twitter for disseminate information to twitter

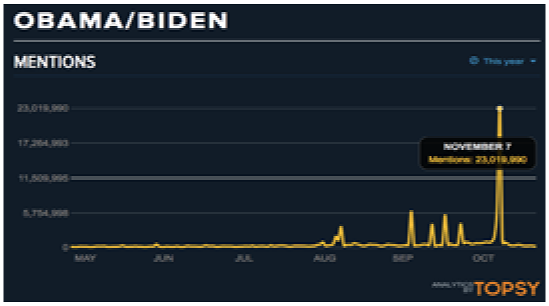
III. LITERATURE REVIEW.

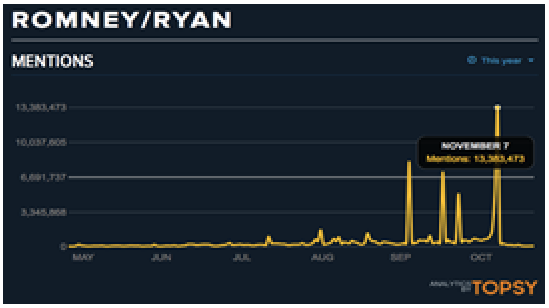
Twitter has been very popular among social networks and a new way of survey and research has been introduced by this social media network. There is lot of different aspect of research regarding twitter in social media, marketing and news media. Hubberman and others [1] studied the interaction of users in twitter and find the sparse network underlying the friends and followers. Java et al [2] review about the structure of community and intentions of different users in twitter. Lot of prior work is done on blog and analyzing the correlation among reviews and blog. Gruhl and others [3] shows how to mine blog and generate the automated queries in order to predict sales of spike. Mishne and Glance[4] correlate movies scores with blog post sentiments. Their Observation for correlation of sentiment are very low and not much sufficient to use for prediction. Tumasjan et al. [5] discussed about the prediction of german elections 2009 using twitter. This paper as clearly two parts, in first part he perform superficial analysis using LIWC (Linguistic Inquiry and Word Count) of tweets. In second part he stated that the number of mentions of parties in tweets actually represent the election results. Moreover he claim that MAE (Mean absolute error) of the prediction on the twitter is more close to the actual polls. Ratkiewicz et al. [6] discussed about the truthy project. This project was to detect political campaign to simulate among the support for a candidate or to spread disinformation. Jungherr et al [7] discussed about the Tumasjan paper that the this was based on choice and its results depend upon time to compute them. Bermingham et al [8] dicuss different aspect of sentiment analysis to predictive analysis. They put their method with the test of 2011 general election of Ireland. Tjong et al [9] discuss about the prediction of 2011 Dutch Senate election using twitter.

IV. Discussion of Related Work.

Prediction of election uses the public opinion on politician or on political party from a particular sample to predict the election result. Sample data or public opinion was collected through questionnaires or by telephone calls which can affect the budget of r prediction or of research. As new technology, web survey provides to do that with low cost.

The text through social media can be effective in prediction. for election. In presidential election, 2008, results can be predicted by the number of supporters of facebook[10]. In German election, 2009, it is found that the 40% of the content that is produced is by the 4% of the all users but it still came close to the election poll survey done by other sources[5]. At the same time, debate is open whether social media can collect the unbiased survey and can predict the election poll in election of britisth columbia,2001, message board didn't mention the actual strength of parties.



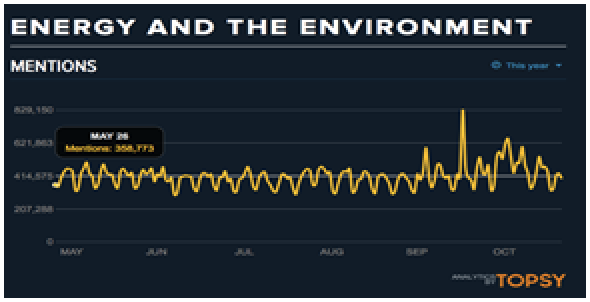


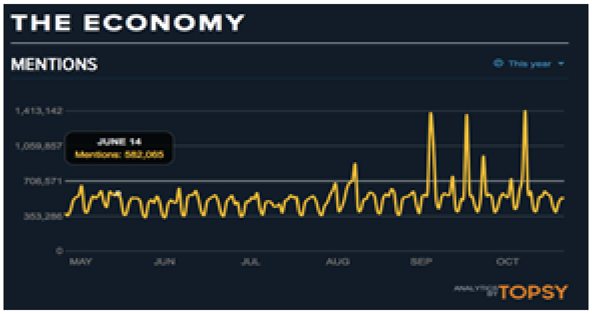
Some other factors can also effect the prediction result as some researchers didn't give the time line when they collected the result set[7]. In web survey, often data set is selected near the election date. If survey made far from the election date it can be meaningless and unfair. It can also be noted that demographics of the society can not be reflected by the social media. In terms of age, 36 percent of citizen, between 18 to 24, 50 percent of citizen between 25 to 34, and 68percent over 35 voted[11]. but on twitter, more than 60% of users are under 24[13]. thus it is said that it could be a biased sampling when collecting data from social media. On other hand, when prediction model is applied on political content generally on social media and Twitter in particular, it is far better than the random classifier to specify the political alignment of users[12][13].

V. Analysis of Related Work.

I have discussed lot of research is continuously in process in this research area. Lot of research has been passed through myself. Http://election.twitter.com has some mention index on base of every day who can win this election poll. They are calculating index by the mention of last name of candidates and it is ranging from 0 to 100.

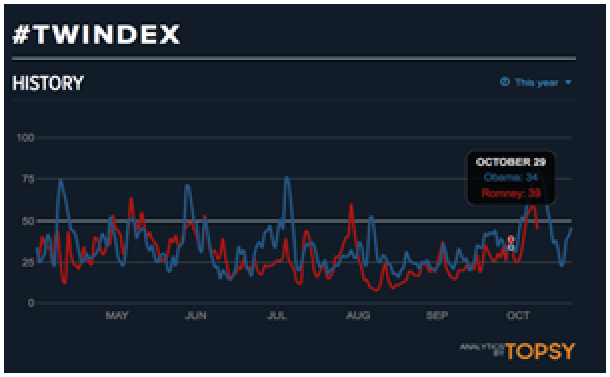
In the above grapgh,it is mention that how many time obama and romney is mention on twitter.





In the above graph, it is mention that what are the subject during the election and what people are discussing over the time. The topics are more but we mention two of them.

In the next graph it show the popularity of obama and biden using index method.



VI. DataSet and Methodolgy.

1) *Sample Dataset.*. I have applied my process to 124052 tweets related to the 124 hashtags through which I get some interesting facts that is mention in below Table 2. it shows that how much keywords are mentions in the tweets in overall total dataset so that we can find some over-view of the data

2) *Twitter DataSet Characteristics.* Twitter data regarding election is collected by hours of crawling feeds of twitter. We started using the crawling the twitter by the hashtag "#USElection". When the tweet regarding hashtags is search, on the result set we again parsed the text of response and and the extract the other hashtags related to it and if it is important hastag like "#obama" or "#romney" we marked it and when again crawling start it will aslo search for this hashtags. Extraction of tweet is over frequent intervals using twitter search API, we have collected timestamp, user name, tweet text for our analysis.Before US election there were three major debate among these two political parties . We try to get tweets using nearby time of these tweets. Some of the important hastags we try to search in tweets are provided in the Table 1.

Hashtags	hastags
USElection	2012PresidentialElection
USA	voteforromney
obama	obamabiden
romney2012	election2012
voteforobama	
romneyryan2012	
republican	
democrat	

In the above table you can see the priority hashtags that is identified for the crawling. This task is one of the most consuming task to identify the important search keywords and get start the more keywords that can be helpful in this reserach. After collection of data we have pre-process the data to implement text-mining on the twitter dataset

3) *Tokenization.* In generating the frequency we have started the tokenization process. Tokenization algorithm generates token that is not only single word but generates the multi-word token. it is started from first word and start making token with second word and then third word until it get "comma", "semi-colon", "fill-stop" or any other common words that is listed as stop word in our file. These stop words are listed on majorly information retrieval sites[13] that are working on term extraction. Belong with the process of token generation it is also start making the frequency of the generated token if the token is not generated it save the token and if the token is already generated it will increment the frequency of that token.

Table 2 : Number of occurrence of hashtags

hashtags	no. of occurrence
Obama	44961
Voteforobama	435
Romney	46724
Election2012	12480
RomneyRyan	14906
romeny2012	5704

4) *Algorithm for Multi-Word Token Generation.* Algorithm that I have applied is given below.

```
Initialize reading document {
        For loop in statement {
                For loop each word
        {
                        If (word ==
        "common")
                                Continue;
                        Generate token
                        If  token exists
                                increment
        frequency of token
                        While (next word
        not stop word){
          Initialize token generation of
```

5) *Analysis of resultset.* In previous section I have discussed the analysis of http://twitter.election.com . We have crawl the twitter search api by the keywords that we mention before and collected the dataset of 124,052 tweets regarding the marked keywords/hashtags. This task is one of the most consuming task to identify the important search keywords and get start the more keywords that can be helpful in this reserach. After collection of data we hsave pre-process the data to implement text-mining on the twitter dataset using Naive Bayes Algorithm. Initially we started with supervised approach by labeling initial text of tweet by "obamawin" , "romneywin" and by "neutral". After initial labeling we apply the model on the remaining tweets. So that it can predict some result. According to the our result we find some interesting fact.

The number of mention of obama and romney is given in Table 7.

hashtags	no. of occurrence
Obama	34893
Romeny	36022

If we look at the table 8 it gives the no of mention of political party republican and democrat

hashtags	no. of occurrence
republican	982
Democrat	3409

\

If we look at user table we can identify the most users are active in the debate of political campaign.

Table : Mentions of active user

hashtags	no. of occurrence
alastyn_wine	370
jm111t	311
NewsDetector	223

In the below table we can check that active users that are tagging obama win

Table : Mentions of users tagging obama win.

hashtags	no. of occurrence
alastyn_mcrider54	24
ZazzleBestSell	19
takecharge188	8

In the below table we can check that active users that are tagging romney win

Table : Mentions of users tagging romney win.

hashtags	no. of occurrence
PhD_Economics	64
mcrider54	24
ZazzleBestSell	19

From the above result set we can find the fact that it is not necessary that every one is using twitter from political use. There can be some users that they are active in this field and this can effect the unbiasedness of the election poll and it is possible that more of the content is been products by less of the user which can effect the unbiasedness of electoral poll.

VII. Processing parameters and assumptions.

I have applied my process on 100,000 tweets related to the 124 hashtags and applied the naive bayes classification algorithms. First of all i have to Tokenize the tweets and then remove the stopwords and implement stemming and generate N-gram algorithm to generate multi-words and then apply Naive bayes classification algorithm to get some model. Then this model is apply to the sample dataset of the twiiter.

VIII. Conclusion& Future Work

Even though, we presented approach to predict the electoral poll using micro-blogging dataset and we found that it is possible to predict the electoral poll but some parameters need to be refined and it can aaffect the prediction of the electoral poll.

1. It is not necessary that social media is used by everyone. It can be biased sample of population. Demographics should be found and prediction has to be corrected on its basis.
2. A very less user is responsible for the lot of political tweeting and therfore thier opinion can drive what is predicted from social media.
3. Simple analysis of sentiments should be avoided for this reserach area of electoral poll as it discuss with double en-tenders , with humor and with sarcasm.

Hence it is listed in challenging problems. and further need the research. hence it is an emerging topic, prediction through social media. Through this we can utilize the wisdom of crowd with low cost and high efficiency.

Acknowledgment

I would like to say thanks to Almighty ALLAH by the grace of whom I am able to complete this research project. Special thanks to my supervisor Mr. Adeel Ahmed for his guidance and support and my friends who help me in this project directly & indirectly.

References

[1] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. First Monday, 14(1), Jan 2009.

[2] Akshay Java, Xiaodan Song,Tim Finin and Belle Tseng.”Why we twitter : understanding microblogging usage and communities”. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65, 2007.

[3] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins. The predictive power of online chatter. SIGKDD Conference on Knowledge Discovery and Data Mining, 2005.

[4] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.

[5] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pages 178–185, 2010.

[6] M Conover, B Gonçalves, J Ratkiewicz, A Flammini, and F Menczer. “Predicting the political alignment of twitter users”. In Proceedings of 3rd IEEE Conference on Social Computing (SocialCom), 2011.

[7] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. “Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to., sander, p.g., & welpe, "predicting elections with twitter: What 140 characters reveal about political sentiment". Social Science Computer Review, 2011.

[8] Adam Bermingham and Alan Smeaton. “On using twitter to monitor political sentiment and predict election results. In Sentiment Analysis where AI meets Psychology”, pages 2–10, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[9] Tjong, E., Sang, K., and Bos, J. “Predicting the 2011 Dutch Senate Election Results with Twitter” . In Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks

[10] C. Williams and G. Gulati, “What is a social network worth? Facebook and vote share in the 2008presidential primaries,” in Annual Meeting of the American Political Science Association, 2008,pp. 1–17.

[11] P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello, “How (Not) To Predict Elections,” in Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Confernece on Social Computing (SocialCom), 2011, pp. 165 – 171.

[12] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj, “Limits of electoral predictions using Twitter,” in Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 2011, pp. 165 – 171.

[13] Sysomos Inc, "An In-Depth Look Inside the Twitter World". http://www.sysomos.com/insidetwitter/. [Accessed Feb 3, 2012].