International Journal of Science Engineering and Advance Technology

Tweet Mining from General Elections of a Country

Muhammad Noman Hayat, Nigar Fida, Haroon Khan, Mohammad Sohail

Department of Computer Science, University of Lahore
Department of Physics, Abdul Wali Khan University Mardan
Software Engineering Department, NUST, Islamabad Pakistan

ABSTRACT-In previous election of a country for its democratic elections in which social media was targeted as a way to convince the social media users to attract their votes, an analysis of BigData on these elections tweets is performed to gather some interesting information. The parties which have been active on social media during elections were identified through hashtag and keyword frequency of occurrences; the top twitter users who influenced the public in election's were identified by identifying the most re tweeted tweets and finally a comprehensive sentiment analysis were done on the big data set to check "Election trolling" where opposition parties insult and attack each other to get an idea of the political maturity of a country twitter users, with a high percentage negative sentiments it was deduced that people of that country lack political maturity.

Keywords: Election trolling, Sentiment Analysis, Twitter Text Mining

INTRODUCTION

Social media is being used in political election campaigns by the respective parties to convince an uncommitted vote for them in the elections. The main idea of the project was to gather the freely available in the general election tweets and then performing text mining on those, the initial task was to gather the tweets the Twitter4J library was integrated with java to gather the streaming tweets, but according to twitter's new change of policy they only provide tweets which are a week older (tweets regarding the query given from past seven days only) than the REST API deployed for
the purpose of testing whether it gets any older tweets than a week but the same thing happened tweets which were older than a week couldn't be acquired.

As twitter only provides its tweets archive or fire hose access to its partners, there was just one way remaining to gather those tweets and that was to get it from some private company which can provide this data. Hence, on request this data was obtained from a private company for research purpose.

I. TWEET STATISTICS

Total Election tweets acquired were 2,635,295 out of these tweets out of these tweets 14,49,412 were retweets. (55%) hence, the actual tweets were 45% this was found by making a dictionary of the data set which had 502,215 unique words than these were listed down alphabetically and frequency wise with highest to lowest frequency. The number of retweets was found out by the word "RT" which refers to retweet. 2,134,588 tweets were in English language. (81%) which was found by the language column "en" value which refers to English Language. Other languages were 'in' Indonesian abbreviation which should have been tagged as Roman Language by tweeter.

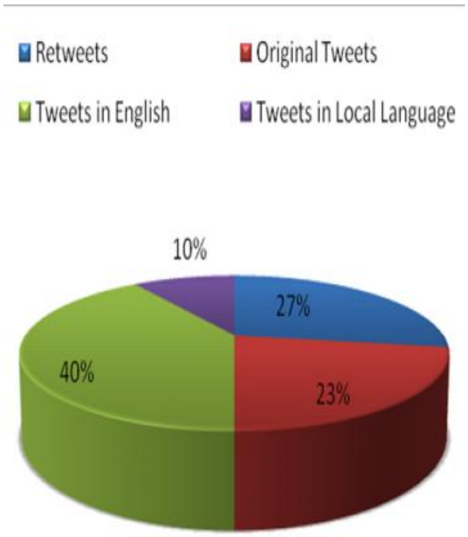

Figure 1: Tweet Statistics

Later the election tweets data was converted into a dictionary and the numbers of words found in data set were 413068. These words were sorted alphabetically and also from highest frequency to the lowest for data cleaning and trend analysis purpose. The two major parties contending against one other are Party-A and Party-B. After the words were sorted by frequency these were arranged in two major contending party groups as under. The keywords and hashtag words given below were marked on a threshold of 10000 frequency; this was decided according to subjective understanding.

TABLE 1. HIGHEST FREQUENCY WORDS IN DATA SET OF PARTY-A & PARTY-B.

Party	Keyword or Hashtag with frequency
Party-A	Party-A (1952684), A-Nickname (290268), A-Candidate (194905), A-Leader (321934) A-LeaderName(146188), A-Slogan (101365), A-LeaderNickname (66445), (59394), A-Family(29648) A-Sign (51684), A-Name's (22863) A-2nd Slogan(31256), A-LeaderLastName(21265), A-3rd Slogan (13994), (11698), support-A (15395) A-Workers (38068),
Party-B	Party-B (395944), B-Leader (48269), B-Fmily (61829), B-ShortName (22344), B-Sign (22564), B-Female(6254), B-Slogan(5241), B-Candidate (1556), B-NickName(23156) Vote-B(2451), B-2nd Slogan(2569), B-Female Candi(25693)

II. ELECTRONICS MEDIA

One of the targets was to identify those twitters whose retweet rate was the highest during the election days, whose are the people who possibly influence other social media users to change or form there opinion other political parties, for this purpose every tweet was tokenized and the respective tweet only column was extracted and saved into a new file on this new file a sting matching algorithm was run (As the data was too large the use of Hashmap was done) the histogram of tweets with their respective occurrence was acquired, the beginning part of tweet after RT is the Twitter Username of the person who's tweet gets most retweeted with the count.

TABLE 2. TWEETS, RE-TWEETS, OF MEDIA AND THEIR RETWEETS

Twitter Coded Username	Retweets in data set	Number of Followers
Mr. X	21, 11, 19	155K
Party-A	20, 30	165K
Party-A	37	75K
Party-A Leader	17	30K
Party-B	23	22K
Party-B	51	1245

III. ANALYSIS AND RESULTS

In paper [1] authors have worked on analyzing the results of on a Country election tweets which were about 50k related sort of analysis was also performed in [2][3][4][5]. Our data set was larger and to check there validity it was to be performed the larger data set we had. Google code [8] for sentiment analysis was taken and modified with respect to our data, Stemming was performed on the testing data and a marked training data set was provided to Base Line Classifier a sentiment analysis classifier which compares each tweet with a list of positive and negative words and calculates the sentimental positive score for a tweet with positive sentiment, and a total negative score for a negative one. The list of positive and negative words was also updated because about 24.6% percent of tweets were in Roman language (tagged by twitter as 'in' language). The words were added according to personal subjective opinion and a list of negative and positive roman language words available at [6]. This analysis is also done on emoticon representation i.e. if a an emoticon appears it will be tagged as a negative sentiment otherwise if a appears it will be tagged as a positive tweet.
The formula for calculation of sentimental score per tweet = Number of negative words per tweet × (-6) + Number of positive words per tweet × (6)

Analysis (a) = (a K)(-6) + (t Q)(6)

IV. CONCLUSION

Twitter is a popular social network where millions of tweets are generated everyday about different events and news. In this the numbers of tweets generated on twitter were around 700 million per day according to twitter blogs. [9] It is a rich and useful source of data because all the tweets are available publically. One can collect the related tweets and perform data mining procedures to acquire interesting information from these tweets. The tweets during this Election were collected which is around 1.5GB. After performing word frequency analysis the most socially accepted parties were identified. Media Hubs were identified whose tweets influenced people towards certain events and parties during election. Sentiment analysis about election trolling was done on the whole data set and a large percentage of negative tweets show that people here consider election trolling the only way of proving their point.

REFERENCES

[1] Younus, Arjumand, et al. "Election trolling: analyzing sentiment in tweets during Pakistan elections 2013." Proceedings of the companion publication of the 23rd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2014.
[2] Chung, Jessica Elan, and Eni Mustafaraj. "Can collective sentiment expressed on twitter predict political elections?" AAAI. 2011.
[3] Tumasjan, Andranik, et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." ICWSM 10 (2010): 178-185.
[4] O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." ICWSM 11 (2010): 122-129.
[5] Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.
[6] Bilingual Roman Urdu sentiment words with weights, NUST Center of Data and Text Engineering and Mining, Sentiment Analysis for Social Network's data in Urdu, https://sites.google.com/a/mcs.edu.pk/codteem/biling_senti
[7] Tool for generating word could, http://www.wordle.net/
[8] Google project on sentiment analysis, https://code.google.com/p/twitter-sentiment-analysis/
[9] Twitter Developer blogs, 2013, https://blogs.twitter.com