PREDICTION OF U.S. ELECTION USING TWITTER DATA: A CASE STUDY

By

ZENIA ARORA

Bachelor of Technology in Computer Science and Engineering

M.D. University

Rohtak, India

2009-2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 2015

PREDICTION OF U.S. ELECTION USING TWITTER DATA: A CASE STUDY

Thesis Approved:

Dr. K.M. George

Thesis Adviser

Dr. Nohpill Park

Dr. Johnson P. Thomas

ACKNOWLEDGEMENTS

Looking into the period of time spent for my thesis work , pursing my Degree of Masters of Science from Computer Science Department, Oklahoma State University has rewarded me to extensive experiences and knowledge in my field. Upon completion of my MS thesis, I take this opportunity to thank people who have been a great help in this period and show my intense gratitude towards them. Foremost I would like to thank my thesis advisor and head of the Computer Science Department, Oklahoma State University, Professor K.M George for his guidance support and encouragement. His constant monitoring made my work progress smoothly and on time. I gained experience from his vast knowledge and suggestions relating to my research work and overall professional development.

I convey my deepest gratitude to my committee members Professors Nohpill Park and Johnson P. Thomas for their guidance and support. I would also like to thank my senior Ashwin Kumar Thandapani Kumarsamy for his extensive discussions and help. I'm also thankful to lab and administrative staff for helping me in many ways. My strength lies in my family and I dedicate my work to them.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: ZENIA ARORA

Date of Degree: JULY, 2015

Title of Study: PREDICTION OF U.S. ELECTION USING TWITTER DATA: A CASE STUDY

Major Field: COMPUTER SCIENCE

Abstract: Social network/media has become popular over the last few years and is moving closer to be an integral part in one's life. With the rise of new social media movement, the analysis of social networking blog contents has become an important tool of big data analytics. Recent research studies on the use of Twitter for predicting political elections have raised many questions as well as interest in using Twitter data for predictive analysis. The overarching objective of our research is to study the capability of Twitter data as an ex-ante indicator of event outcomes. The 2014 US midterm election has been chosen as the event for this study. This work analyses both pre-poll and post-poll data from Twitter related to 2014 midterm elections in U.S. Relevant tweets are extracted from the tweet stream with the help of a Map-Reduce Program in a Hadoop system by specifying appropriate keywords configuration for running Apache Flume. This data are classified into four groups using 'Democrat' and 'Republican' as the division criteria. Two time-series of sentiments (positive and negative) are constructed for each group. Several statistics are also compiled from each group of tweets and used as predictive indicators. Original tweet count, retweet count, and user count in each group are some of the statistics compiled. All the statistics favor the Republican party to win which actually was the outcome of the election.

Our research consists of two parts. The first part is prediction of election results and the second part is modeling sentiment before and after the election. We used Hidden Markov Model as a tool for both parts. The hidden states of the model were used as sentiment indicators and state changes were interpreted as sentiment changes. The results of the HMM agreed with the actual outcomes. Our study provides support for the argument that Twitter data can be considered as a reliable predictor of events.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION..1

1.1 Research Objective and Questions..1
1.2 Scientific Approach 2
1.3 Present Work..3
1.4 Thesis of Outline..5

II. REVIEW OF LITERATURE..6

2.1 Importance of Opinion Polls in Election 6
2.2 Twitter Review...7
2.3 Election Prediction using Twitter 8
2.4 Sentiment Analysis 9
2.5 Data and Information Extraction From Microblog 11

III. METHODOLOGY 13

3.1 Procedure Adopted for Extracting Data...13
3.2 Description of Tools for Obtaining Data 18
3.2.1 Apache Hadoop...18
3.2.2 Apache Flume 21
3.4 Hidden Markov model..23
3.3 JSON Format 24

Chapter Page

IV. FINDINGS.....32

4.1 Tweet Count.....33
4.2 Retweets.....39
4.2.1 Comparison of Original Tweets and Retweets- Pre-poll and Post-poll...40
4.2.2. Retweet Frequency.....49
4.3 Users and Tweets.....56
4.3.1 Comparison of Total Distinct Users.....56
4.3.2 Number of Top10 and Bottom 10 Users.....58
4.3.3 Average Number of Tweets Done Per User.....67
4.4 Sentiment Analysis.....68
4.5 Correlation of Tweets/Retweets to Election Result.....73

V. CONCLUSION.....75

5.1 Prediction and Sentiment Change Based on Counts.....75
5.2 HMM Analysis.....76
5.3 Future work.....76

REFERENCES.....77

APPENDICES.....82

1. Keywords for Second Level of Filtration.....82
2. Negative Keywords and Phrases for Third Level of Filtration.....84
3. Positive Keywords and Phrases for Fourth Level of Filtration.....86
4. Steps for Installing Hadoop and Flume.....89
5. List of Initial Probabilities assumed for each category.....90

LIST OF TABLES

Table Page

1. Keywords for initial filter 14
2. Predictive Ratio for Total – Tweets and Retweets..73
3. Prediction Ratio for Positive and Negative – Tweets and Retweets.....................74
4. Keywords for Second Level of Filtration 82
5. Negative Keywords and Phrases for Third Level of Filtration.............................84
6. Positive Keywords and Phrases for Fourth Level of Filtration.............................86
7. Steps for Installing Hadoop and Flume 89
8. List of Initial Probabilities assumed for each category...90

LIST OF FIGURES

Figure Page

1. Block diagram shows how the relevant tweets were extracted from new data 16
2. Map- Reduce Programming Structure 18
3. Hadoop File System and Data Communication Architecture..............................20
4. Basic Architecture on Streaming Data Flow 22
5. Data-Flow of Analysis of Twitter Stream Data 23
6. Comparison of positive tweets –DEM, REP and DNR 34
7. Comparison of negative tweets – DEM, REP and DNR 34
8. Pre-poll Positive tweets – DEM, REP and DNR 35
9. Pre-poll Negative tweets – DEM, REP and DNR...36
10. Post-poll Positive tweets – DEM, REP and DNR..37
11. Post-poll Negative tweets – DEM, REP and DNR 37
12. Pre-poll Positive and Negative tweets – NDR...38
13. Post-poll Positive and Negative tweets –NDR 39
14. Total number of positive tweets that are retweeted 39
15. Total number of negative tweets that are retweeted 40
16(a). Pre-poll Positive Tweets – Original Tweets and Retweets for REP 41
16(b). Pre-poll Negative Tweets – Original Tweets and Retweets for REP............42
17(a). Pre-poll Positive Tweets – Original Tweets and Retweets for DEM 42
17(b). Pre-poll Negative Tweets – Original Tweets and Retweets for DEM 43
18(a). Pre-poll Positive Tweets – Original Tweets and Retweets for DNR 43
18(b). Pre-poll Negative Tweets – Original Tweets and Retweets for DNR...........44
19(a). Pre-poll Positive Tweets - Original Tweets and Retweets for NDR 44
19(b). Pre-poll Negative Tweets – Original Tweets and Retweets for NDR...........45
20(a). Post-poll Positive Tweets – Original Tweets and Retweets for REP 45
20(b). Post-poll Negative Tweets – Original Tweets and Retweets for REP 46
21(a). Post-poll Positive Tweets – Original Tweets and Retweets for DEM 46
21(b). Post-poll Negative Tweets – Original Tweets and Retweets for DEM..........47
22(a). Post-poll Positive Tweets – Original Tweets and Retweets for DNR...........47
22(b). Post-poll Negative Tweets – Original Tweets and Retweets for DNR..........48
23(a). Post-poll Positive Tweets – Original Tweets and Retweets for NDR...........48
23(b). Post-poll Negative Tweets – Original Tweets and Retweets for NDR..........49
24(a). Top 10 positive pre-poll retweets for DEM 50
24(b). Top 10 positive pre-poll retweets for REP..50
25. Top 10 positive retweets: DEM and REP – Pre-poll 50
26(a). Top 10 positive pre-poll retweets for DNR...51

Figure Page

26(b). Top 10 negative pre-poll retweets for DNR..51
27(a). Top 10 negative retweets for DEM 51
27(b). Top 10 negative retweets for REP...51
28. Top 10 negative retweets between DEM and REP – Pre-poll............................52
29(a). Bottom 10 positive pre-poll retweets: DEM, REP and DNR.........................52
29(b). Bottom 10 negative pre-poll retweets: DEM, REP and DNR........................52
30(a). Top 10 positive post-poll retweets for DEM...53
30(b). Top 10 positive post-poll retweets for REP 53
31. Top 10 positive retweets: DEM and REP - Post-poll 53
32(a). Top 10 positive post-poll retweets for DNR 54
32(b). Top 10 negative post-poll retweets for DNR 54
33(a). Top 10 negative post-poll retweets for DEM..54
33(b). Top 10 negative post-poll retweets for REP 54
34. Top 10 negative retweets: DEM and REP – Post-poll..55
35(a). Bottom 10 positive retweets: DEM, REP and DNR – Post-poll 55
35(b). Bottom 10 positive retweets: DEM, REP and DNR – Post-poll....................55
36. Comparison: Total useful tweets and Retweets – Pre-poll and Post-poll...........56
37. Total distinct users for Positive Tweets..57
38. Total distinct users for Negative Tweets 58
39(a). Top 10 users for pre-poll positive tweets – DEM 58
39(b). Top 10 users for pre-poll positive tweets – REP..58
39(c). Top 10 users for pre-poll positive tweets – DNR...59
40. Top 10 users: DEM, REP and DNR - Pre-poll positive tweets 60
41. Bottom 10 users: DEM, REP and DNR- Pre-poll positive tweets......................60
42(a). Top 10 users for pre-poll negative tweets – DEM 61
42(b). Top 10 users for pre-poll negative tweets – REP...61
42(c). Top 10 users for pre-poll negative tweets – DNR...61
43. Top 10 users: DEM, REP and DNR - Pre-poll negative tweets 62
44. Bottom 10 users: DEM, REP and DNR- Pre-poll negative tweets.....................63
45(a). Top 10 users for post-poll positive tweets – DEM...63
45(b). Top 10 users for post-poll positive tweets – REP 63
45(c). Top 10 users for post-poll positive tweets – DNR 64
46. Top 10 users: DEM, REP and DNR- Post-poll positive tweets..........................64
47. Bottom 10 users: DEM, REP and DNR- Post-poll positive tweets 65
48(a). Top 10 users for post-poll negative tweets – DEM..65
48(b). Top 10 users for post- poll negative tweets – REP 65
48(c). Top 10 users for post-poll negative tweets – DNR 66
49. Top 10 users: DEM, REP and DNR –Post-poll negative tweets 66
50. Bottom 10 users: DEM, REP and DNR- Post-poll negative tweets 67
51(a). Average positive tweets done..67
51(b). Average negative tweets done...67
52. Total number of common distinct users...68
53(a). Sentiment analysis for DEM positive tweets..69
53(b). States for pre-poll and post-poll – DEM positive tweets 69

Figure Page

54(a). Sentiment analysis for DEM negative tweets..69
54(b). States for pre-poll and post-poll – DEM negative tweets 69
55(a). Sentiment analysis for REP positive tweets...70
55(b). States for pre-poll and post-poll - REP positive tweets 70
56(a). Sentiment analysis for REP negative tweets 70
56(b). States for pre-poll and post-poll – REP negative tweets................................70
57(a). Sentiment analysis for DNR positive tweets..71
57(b). States for pre-poll and post-poll – DNR positive tweets................................71
58(a). Sentiment analysis for DNR negative tweets 71
58(b). States for pre-poll and post-poll – DNR negative tweets...............................71
59(a). Sentiment analysis for NDR positive tweets..72
60(a). States for pre-poll and post-poll – NDR positive tweets................................72
60(b). Sentiment analysis for NDR negative tweets...72

CHAPTER I

INTRODUCTION

Social media and microblog services such as Twitter have grown popular in the recent years and are large sources of information [4, 5]. Twitter allows users to post messages (upto140 characters) that are publically visible through Internet. One could track people's opinion about each candidate by capturing tweets and analyzing the sentiments behind those tweets. In this manner, earlier researchers [2] were able to predict election results through Twitter**.**

1.1 RESEARCH OBJECTIVES AND QUESTIONS

"Digital democracy is here. We no longer passively watch our leaders on television and register our opinions on Election Day. Modern politics happens when somebody comments on Twitter or links to a campaign through Facebook" [23]. The quote captures the significant role of microblogs in elections. Opinion polls/surveys play an important role in the functioning of democracy. Polls provide information to interested citizens. Political parties use polls to measure the effects of campaign strategies and make adjustments accordingly. Opinion polls existed since early 1900s and has evolved since [1]. Using social media for political change and predicting election results have been the topic of several papers [3] At least one author have claimed that it is not possible to predict election results with Twitter data [2]. Researchers have become increasingly interested in the question "Whether success on Twitter can be taken as the indicator for electoral success when there is ever increasing use of Twitter by parties, politicians, general public during elections and political campaigns" [4, 6]. There are conflicting claims regarding predicting elections results by using twitter. A strong correlation of Twitter with presidential election has been reported by O'Connor et al [7] and Tumasjan el al [8]. On the other side many claimed Twitter to be a poor electoral predictor [9].

1.2 SCIENTIFIC APPROACH

Bollen et al [10], was among the first few who worked on the application of mood analysis of Twitter data. O' Connor el al [11] worked on feasibility of using Twitter data as a substitute for traditional polls. Using this method for US 2008 Presidential approval poll found a strong correlation with Twitter sentiment data. Mustafaraj et al [12] introduced the concept of "Twitter bomb", the concept of use of fake accounts in Twitter to spread disinformation by 'bombing' targeted users who, in turn would re-tweet. Mislove et al [13] analyzed Twitter user data and compared to US population, using the parameters geography, gender and race. It was found that highly populated countries are over represented on Twitter, users are predominantly male and Twitter is a non-random sample with regards to race/ethnicity. Castillo et al [14] was successful in implementing a method to separate credible from non-credible tweets.

US Presidential election 2012 showed a tense battle between two key candidates. The campaign lasted several months and affected the sentiment and twitter. The campaign of President Obama in 2008 demonstrated the power and reach of social media. Political analysts [15] attributed the success to the active and effective use of media to engage voters mainly the younger generation but less importance was given to elderly politicians. However lack of recognition for "no online population influences" on political landscape is still an important aspect to deal with [16]. Gayo - Avello et al [2] listed the core criticism of electoral prediction stating that effect of incumbent is not measured by the researchers, no unified approach to modeling of tweets and sentiments analysis is available. Furthermore, demographic information should be used for analysis. Livne et al [17] used the study of twitter for House and Senate candidates during the midterm (2010) elections in US. Graph and text mining techniques were used to analyze differences between Democrats, Republicans and Tea Party candidates and suggested a novel use of model to predict the results.

According to the Asur et al [19] over 80% of Americans use at least one social network and spend 23% of their online time of social networks. With over 140 million active users, generating 340 million tweets per day [20] Twitter has become an important instrument for prediction and opinions in America. The challenge remains that all the users are treated equally. However, recent works [21, 18] show that social media users from different groups have different tweeting behavior patterns. The effect of different tweeting and generating behavior on prediction of election results are yet to be looked into. Recent work by Manish et al [22] devised a technique to include relevant and filter irrelevant tweets based on predefined set of keywords. This work claimed to approach success in predicting the winner of all three presidential elections in Latin America during 2013. Viewing all the previous literature, there is an indication that it is very important to look in to the following facts:

- Not everybody is using twitter.
- Not every twitter is tweeting politics.
- All twitters are not true and simplistic sentiments analysis methods are not enough for analysis.

1.3 PRESENT WORK

The present work is done to analyze the effectiveness of traditional time series based forecasting techniques applied to twitter data in predicting 2014 US midterm elections. The approach is similar to what has been used for Swedish General Election campaign [18]. By using the social network twitter, a large data set is collected with the help of Hadoop based tools.

The methodology used for this experiment involves, fetching data from twitter using Flume. Configuring flume configuration file to catch tweets which contains keywords like US elections, senate, democrats, republicans, senate, midterm elections, Obama, US president, etc. The data fetched will be stored on Hadoop Distributed File System. This data may contain a lot of unwanted tweets to which we refer as noise. To eliminate these tweets we used Map-Reduce program. The map part of program selected only those tweets, which contains keywords "us elections", "uselections", "democrats", "republicans", "senate", "midterm elections", and "midtermelections". The Reduce part of program classified the filtered tweets into the following four groups:

- Tweets talking only about Republicans.
- Tweets talking only about Democrats.
- Tweets talking about both Republicans and Democrats.
- Tweets talking about none of them.

After the filtration, a program was written in java language to segregate the negative and the positive tweets for each of the above categories using the negative phrases and negative keywords. To make sure that the positive and negative tweets are separated accurately, second level of filtration was done using positive phrases and positive keywords. For each of the four categories i.e. Democrats, Republicans, Both, and None, a file of positive tweets and negative tweets was created.

Then time series was formed for each of the groups for the time during which the tweets were captured. To predict the outcome of the tweets the predictive models Hidden Markov Model (HMM) was used. The final step was to analyze the predicted value against the actual outcome and a predictive capability of the model was decided. User interest in the parties before and after the election are also analyzed.

1.4 THESIS OUTLINE

In the next chapter the information of background review and literature survey is presented in details. The chapter starts with importance of opinion polls in election and Twitter review, followed by several related works and their analyses. The chapter ends with description of the extraction of data and information from microblog.

In the third chapter, the methodology followed in this work to predict the election is discussed in details. The procedure adopted for extraction of data is framed. Description of tools used for obtaining data like Apache Hadoop, Apache Flume are looked into. JSON format is also described in this chapter.

The fourth chapter will explain the results, interpretations and analysis. This leads to the fifth chapter where contributors and limitations of this work will be discussed. Suggestions and scope for future research are presented.

CHAPTER II

REVIEW OF LITERATURE

This chapter explores background information that are the basis for this research work and the tools that are made use of in conducting the work. This chapter is subdivided into:

- Importance of opinion polls in elections
- Twitter Review
- Election prediction using Twitter
- Sentiment analysis
- Extraction of data and information from Microblog.

2.1 IMPORTANCE OF OPINION POLLS IN ELECTIONS

The history of opinion polls is as old as 1824[1] were it was used to predict the results of US presidential election by taking information from scattered offices and public meetings. Scientific methods of predicting poll results started only in the beginning of 1940’s when Gallup Poll first released a survey of polling results in presidential pre-election. Lewis-Beck [24] summarized, that study of statistical model for pre-election prediction started to appear around 1980 in the US, UK and France. Since then lot of research has been conducted [25] to find ways to predict the results of election polls. To resolve these issues, which involve accuracy and high cost, study of possibility of using data from social media as data source to predict outcome of election has gained lot of importance.

2.2 TWITTER REVIEW

Twitter is a social media platform/micro blog. The main form of communication on Twitter is a tweet, a short string of characters. If one person is a follower of a celebrity, this means that celebrity's tweets are immediately visible to the follower from the individual's homepage. Celebrity is one of the individual's followees. Twitter users can also send messages directed specifically at other users, called mentions, using the syntax "@ [username]" anywhere in the tweet. Mentions can be sent to anyone on Twitter. Common conventions regarding tweet contents are also to be looked into. If a user enjoys someone else tweet and wishes to share it with his own followers, she/he can re-tweet it, thus sending the same message as one of her/his own tweets. Re-tweet often contains an acknowledgement of the original poster using "RT@ [username]" or "via @ [username]" syntax. Users can also characterize tweets using hashtag denoted by syntax "# [word]". Hashtags are generally contained at the end of a tweet and indicate the general topic of tweet such as "#SOTU" (State of Union). Users can also share links in their tweets using any number of link-shortening services, which create links that re-direct to longer URL.

Researchers at Microsoft [26] in their research in 2010 investigated frequency of these types of tweets using a random sample of tweets. It was found 36% of tweets were mentions, 5% of tweets contain hashtags, 3% of tweets where re-tweets and 22% of tweets contained URL. Re-tweets were more likely to contain hashtags (18% of re-tweets) or a URL (52% re-tweet).

Twitter is unique in many ways. It is found that only 22% of Twitter relationships are mutual. Twitter does not impose a limit on the number of followers. This relationship can be modelled as a directed graph. Both in degrees (number of followers) and out degrees (number of followees) of Twitter follow a power-law-distribution [27]. Active Twitter users commonly follow celebrities and media personalities with whom they are not personally acquainted. An individual who has mentioned at least two tweets is a “friend”. Friend-to-followees ratio on Twitter is 0.013 and the median is 0.04 [28]. Number of friends gradually saturates quickly as the number of followees rises [28]. It is seen that users often connect with individuals with little intention of active communication with them. Indirectly many use Twitter, as a way for passive interaction by sending other’s updates. Further it may be noted that number of re-tweets that any tweet receives appears also to follow a power law [29], concluding that most tweets receive very little attention but handful receive a very significant attention.

2.3 ELECTION PREDICTION USING TWITTER

Being able to make predictions based on publicly available data would have many benefits [30, 31, 32, 33] especially in politics. There have been reports on Twitter ability to predict with accuracy the voting results in recent 2009 German elections [8] and in the 2010 US Congressional elections [34].

The word ‘prediction’ means foreseeing the outcome of events that have not yet occurred. There is relatively high amount of hype regarding feasibility of predicting electoral results using social media. The hype is fueled by traditional media and blogs. Shortly after the recent 2010 election in US, people argued that Twitter is not a reliable predictor [35] to those claiming opposite [36]. The degree of accuracy of these predictors was usually assessed in terms of percentage of correctly guessed electoral races, for example for winners 74% for US House and 81% for US senate races were predicted [37] correctly. Tumasjun et al [8] focuses directly whether Twitter can serve as a predictor of electoral results. He wrote “the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its prediction power even comes close to traditional election polls”. Mean average error of only 1.65% was reported. They found that “co-occurrence of political party mentions accurately reflected close political positions between political parties and plausible coalitions”.

A. Livne et al [34] used the tweets sent by electoral candidates not the general public and reported success in “building a model that predicts whether a candidate will win or lose with accuracy 88%.”

Mustafaraj et al [21] inferred that “prediction theory should be an algorithm with carefully pre-determined parameters and data analysis should be aware of the difference between social media data and natural phenomenon data”. Moreover they emphasized that one should establish a sampling method comparable to the ones used by professional pollsters.

2.4 SENTIMENT ANALYSIS

There appears to be three research areas emerging in terms of using online sentiments [38]. The first is event monitoring where the aim is to monitor reactionary content in social media during an event like a speech or debate. Shamma et al [38] examined variety of aspect of debate modeling using Twitter. Twitter proved to be effective source of data for identifying important topics and related public reactions. The second area, which has received attention, is modeling continuous sentiment functions for predicting other real-world values. Bollen et al [39], Zhang et al [40], O Connor et al [41] studied the emotive sentiments (mood states, emotions) rather than polar sentiments (positively, negatively). The third related area is result forecasting. In result forecasting, it is the final result, which is used to judge the accuracy of particular forecasting measures. Asur and Huberman [42] find Twitter volumes to be a strong predictor and sentiments to be useful, but a weaker predictor. They propose a general model for “linear regression social media prediction”. Tumasjan et al [43] found that volume on Twitter accurately reflected the distribution of votes. Gayo-Avello [44] argued that one should not be accepting prediction about events using social media data as a “black box”. Jungherr et al [57] echoed, “Prediction using social media analytics are frequently contingent on arbitrary experimental variables”.

Twitter has received substantial attention in research all over the world. Bernado et al [45], Devin Gaffney [46] and Anders Olof Larson [47] used graphical methods for mapping and categorized the tweets. ‘Alternate Graph Ranking Algorithm’ was tested by Ghosh et al [48]. Topological relationships within a graph are important for predictions [49]. The Twitter ‘graph’ consists of the collection of nodes (representing users) and edges (representing following relationships). ‘Histogram based analysis of users and tweets’, ‘Network based analysis of re-tweet called Re -tweets influence map’ and ‘Simple language based analysis’ are frequently used for prediction [46]. Devin Gaffney [46] studied Twitter used in Iran election while Jennifer Golbeck et al [49] analyzed tweets from US Congress. Hallvard Moe et al [50] focused on Norwegian blogosphere and Axel Bruns et al [51] reported on Australian blogosphere. Further Anders Olof Larson et al [47] studied Swedish Election campaign in the year 2010. They all studied and looked into the insight of specific part of Twittershpere and focused on need of empirical research that can track the networks qualitatively and quantitatively. They further looked into the key methodological methods of predictions.

Kwak et al [27] investigated the relationship between the simple in-degree of a Twitter user and his or her influence and concluded that there is a substantial difference between the list of the most followed individuals and most re-tweeted individuals.

Studying influence pattern can help to understand trends and innovations, which in turn influence campaigners to look into effective measures. The analysis of the influence of Twitter users is done by three measures (1) indegree (2) retweets and (3) mentions. Large amount of data was collected from Twitter and companied using these measures by Meeyoung Cha et al [52]. They found that popular users who have high in-degree are not necessary influential in terms retweets or mentions.

2.5 DATA AND INFORMATION EXTRACTION FROM MICROBLOG

Since the birth of twitter in October 2006, it has a potentially large information base in the field of data mining. Information detection and extracting useful information is a challenging task. Tweets are generally noisy, ungrammatical and full of abbreviations, symbols, and misspelling. They cannot be used as it is. Information extraction is an automatic extraction process to generate structured data from a collection of unstructured or semi-structured documents, for post processing. The challenging task involves entity extraction (Entities are referred by names such as people, organization or location). According to Guo J et al [53] "named entities occur in about 71% of search queries". Wen Hua et al [54] mentioned that there are three types of sources in microblogs, from which one can mine useful information: 1) the metadata containing user profiles or tweets. 2) Content of tweets and 3) Network structure. All three types of sources are related to each other. Hua et al [54] specified, "We are attempting to conduct friend recommendation based on user's co-location patterns. We consider it to be a trending topic in near future, with the rising prevalence of location –based services". Kalina et al [55] concluded that the structure of the tweets makes them difficult to interpret and exhibit much more language variations. To combat these problems, research was focused on microblog-specific information extraction algorithms [56,57]. Microtext normalization is a way of removing some of linguistic noise [58, 59,60].

Ritter et al [58] took a "pipeline approach performing first tokenization and tagging before using topic modes to final named entities ". Liu [62] proposed a "gradient-descent graphs" based method for doing joint text normalization and recognition. Kalina et al [55] have included evaluation of Twit IE, a twitter adapted version. Twit IE focuses on named entity recognition. Min Peng et al [63] proposed a high quantity information extraction framework based on the idea of multiple features like comment number, forward number, URL, textual content and follower number. Min Peng used three steps, First: Construction of k-dimensions feature matrix as an extraction basis matrix. Second: Transformation of k-dimensions feature into time frequency domain. Third: Estimation of each feature's contribution to information quality based on EM algorithm.

Hua Zhao et al [64] worked on to find ways to extract the keywords effectively from a single microblog text. They found a method based on the combination of multiple features. "The method involves weight feature based on graph model, the statistical feature based on the semantic and the location of the word".

CHAPTER III

METHODOLOGY

3.1 PROCEDURE ADOPTED FOR EXTRACTING DATA

In this section we outline the process followed to collect data for our experiment. The data source is tweets from Twitter. As there was no research on the best time to collect data for the experiment, we used our intuition for the choice of the period. The data collection period was chosen to be around polling date, when it is anticipated that, people discussed maximum politics and tweeted regarding it. This decision was taken with the assumption that more Twitter users will be active close to the polling event. Hence, the data for my experiment was obtained two weeks prior to the Mid-term Election (4th Nov 2014). Also three months after the election, Twitter data was collected to study sentiment change after the election. The tool used for fetching the data was Apache Flume, as it is efficient and flexible in moving large amount of streaming data. Apache Flume is a tool designed and implemented by Apache Software Foundation. The purpose of Flume software is to collect, aggregates, and move large log data files from many sources to a central storage in HDFS [65].

Flume Configuration File requires key keywords as search criteria to be used for data extraction from all tweets. The keywords used for this purpose are listed in Table 1. In order to determine the keywords, we followed a two-level classification scheme that first identifies the category of tweets and then the keywords describing the category.

Table 1. Keywords for initial filter.

Category	**Keywords**
Event	Elections, Exit poll, Democracy, 4th November
Terms Used for Election	Ballot box, Vote, Ballot, Referendum, Public vote, Plebiscite
Office	Senates, President, Leader
Public Opinion	Popular, Politics, Leadership
Election Related	Pollster, Governance, Government

Table1 shows the list of keywords for initial filtration. The above keywords were added to the Flume Configuration File, which could collect raw data related to General Elections. The obtained raw data was in JSON format [71][72], which is an open standard format. The data was then stored in Hadoop Distributed File System. Hadoop Distributed File System is designed to store a very large data reliably and to stream those data sets with higher bandwidth to user application. The pre-poll raw data size was 50 MB and the post-poll raw data size was 34 MB.

The raw data obtained from Twitter using the keywords listed, contained a significant unrelated tweets or noise. As our experiment focused on the US mid-term elections, tweets not related to this particular event needed to be filtered out. The task of segregating the relevant tweets was done with the help of a Map-Reduce Program. We used the Map-Reduce program to extract only tweets related our research topic from the raw data set that was first collected. We did a second level of filtering by classifying the dataset into four subsets using different keywords. The keywords were chosen using the name of the parties, e.g. democrats and republicans and also name of the candidates e.g. Sid Hill, Chris Coons, Chris Coons Tom Janich. The detailed list of keywords is given in **APPENDIX 1**. The four subsets are:

- REP – tweets that contain reference to Republicans but there are no references to Democrats;
- DEM - tweets that contain reference Democrats to but there are no references to Republicans;
- DNR **-** tweets talking about Republicans and Democrats.
- NDR **-** tweets that have no explicit reference to either party.

The midterm elections are generally influenced by political parties and several candidates, not just one candidate as during the presidential election. So, the motivation for the above mentioned classification is to analyze the strength of parties and candidates. It is obvious that the tweets about midterm election must fall into one of the classes. We use the different classes to perform sentiment analysis and detect possible changes in sentiment following the election. The block diagram in Figure 1 shows the process used for the classification of the tweets.

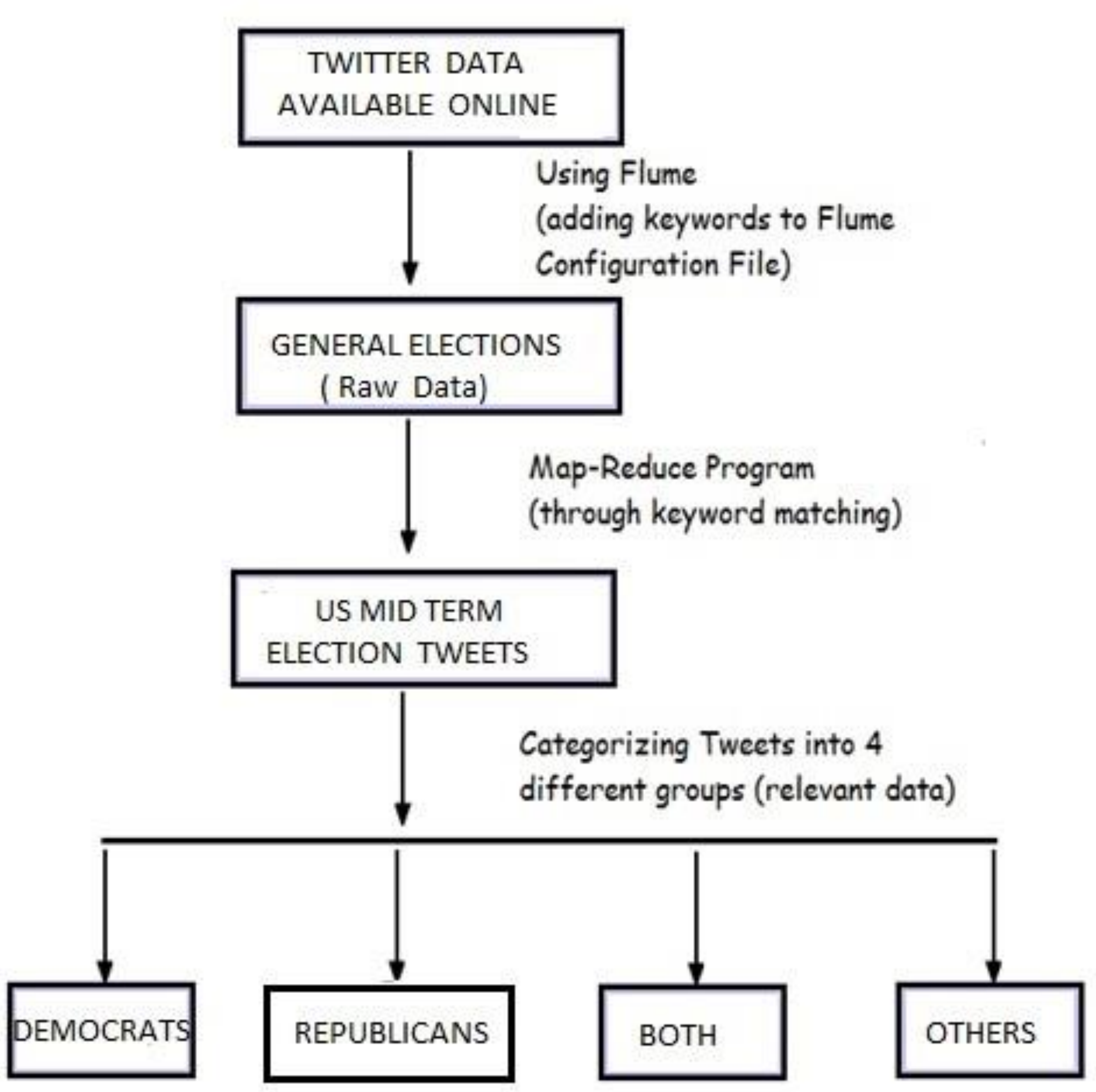

TWITTER DATA AVAILABLE ONLINE
Using Flume (adding keywords to Flume Configuration File)
GENERAL ELECTIONS (Raw Data)
Map-Reduce Program (through keyword matching)
US MID TERM ELECTION TWEETS
Categorizing Tweets into 4 different groups (relevant data)
DEMOCRATS
REPUBLICANS
BOTH
OTHERS

Figure 1. Block diagram shows how the relevant tweets were extracted from raw.

After the filtration, the negative and the positive tweets were to be separated. A program in Java language was written to separate out positive and negative tweets by using negative keywords and phrases given in **APPENDIX 2** for second level filtration. To make sure that the positive and negative tweets are segregated, third level of filtration was done using positive keywords and phrases, given in **APPENDIX 3**.

For each of the four categories i.e. REP, DEM, DNR and NDR, files of positive tweets and negative tweets were created. Therefore we get total eight files: four files of positive tweets and four files of negative tweets.

SQLite, software is used to fetch the data, which is to be analyzed. For each category REP, DEM, DNR and NDR (positive and negative), database tables were created. SQL queries were used to obtain the following:

- Total tweets per day for pre-poll and post-poll
- Total retweets per day for pre-poll and post-poll
- Top 10 positive retweets for pre-poll and post-poll
- Bottom 10 negative retweets for pre-poll and post-poll
- Distinct users for pre-poll and post-poll
- Top 10 positive users for pre-poll and post-poll
- Bottom 10 positive users for pre-poll and post-poll
- Average tweets done by user - pre-poll and post-poll
- Total number of common distinct users.

We have also tried to look into the correlation between tweets/ retweets to actual election results.

Hidden Markov Model (HMM) has been used to analyze the sentiment. Time Series for different categories were obtained. Initial probabilities for the transition vector and emission vector were assumed for each of four categories. Suitable probabilities were assumed taking into consideration that the sum should not exceed one. Also different ranges in a particular data were made and sequences of points were obtained. Sequence of points for pre-poll, probabilities of transition vector and emission vectors were used as parameters to train the HMM model that provide the value of the states for the pre-poll. Further, sequence of points for post-poll was fitted into trained HMM model to figure out the states and was then compared to the states of the pre-poll. **APPENDIX 5** provides all the initial probabilities used for HMM model.

3.2 DESCRIPTION OF TOOLS USED FOR OBTAINING DATA

In this section, we describe the Apache tools that are being used in this research.

3.2.1 APACHE HADOOP

"One of the largest technological challenges in software systems research today is to provide mechanisms for storage, manipulation, and information retrieval on large amounts of data." [65]. Apache Hadoop is an open source software system that supports the storage, retrieval and processing of "big data". Big data refers to data that exhibit the characteristics of volume, velocity, or variety. Generally speaking, big data is unstructured, arrives in high speed, high volume, and in several formats. Hadoop implementation is based on support for the Map-Reduce programming model for massively parallel systems popularized by Google [66]. The architecture of the Map-Reduce programming model is shown in Figure2.

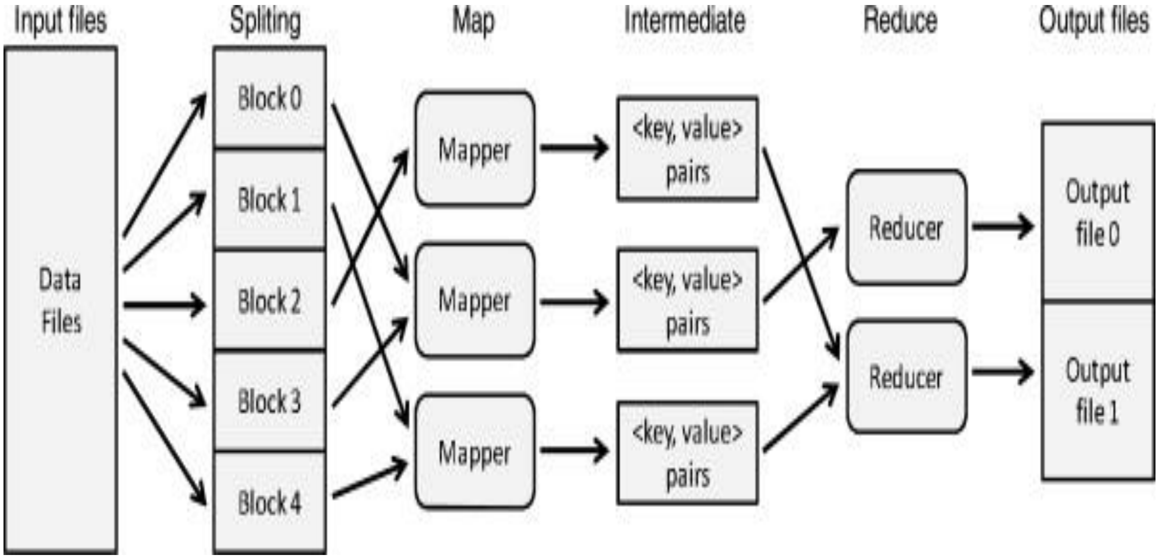


Figure 2. Map-Reduce Programming Structure [67].

As can be seen in Figure 2, data is split into blocks and stored in a distributed manner. Hadoop supports a distributed file system known as HDFS, which is derived from the Google File System. The data communication structure is shown in Figure2. A typical organization of a Hadoop system will consist of several processing units (or virtual units) called nodes. The HDFS file system has master/slave architecture. One distinguished node called the NameNode is designated as the gatekeeper and manager of the entire HDFS file system. Other nodes are called DataNodes. Data are stored as files. Files are split into blocks. Blocks are distributed among the DataNodes. Block distribution is handled by the Hadoop system. Block information is stored in the NameNode. A client wanting to store or retrieve data gets the block information from the name node and directly accesses the block from the data node.

The Apache Hadoop system is broadly classified into two components, namely the MapReduce Engine and the HDFS Cluster. The MapReduce Engine consists of a JobTracker running on the NameNode and TaskTrackers running on the DataNode. The HDFS Cluster consists of the physical nodes, NameNode and DataNodes. The following classes are components in the Hadoop system:

- Hadoop master nodes: In Hadoop interface, a TaskScheduler and class Job In Progress play the role of masternode.
- Hadoop slave nodes: Class TaskTracker is a process on daemon process on every slave node. It receives task execution commands from master node.
- User Define Map and Reduce function Class: Class MapTask and Reduce Task are user defined Mapper and Reducer. Once a TaskTracker gets task execution commands from the TaskScheduler, it starts MapTask or ReduceTask.

The Distributed HDFS file system, which is similar to Google File System Hbase, is efficient in handling semi-structured data. The Hadoop file system architecture is show in Figure 3. Files are split into blocks and replicated among many DataNodes. NameNode keeps the metadata necessary to store and retrieve data blocks. Data storage and manipulation includes storage, replication, indexing and random access queries, etc.

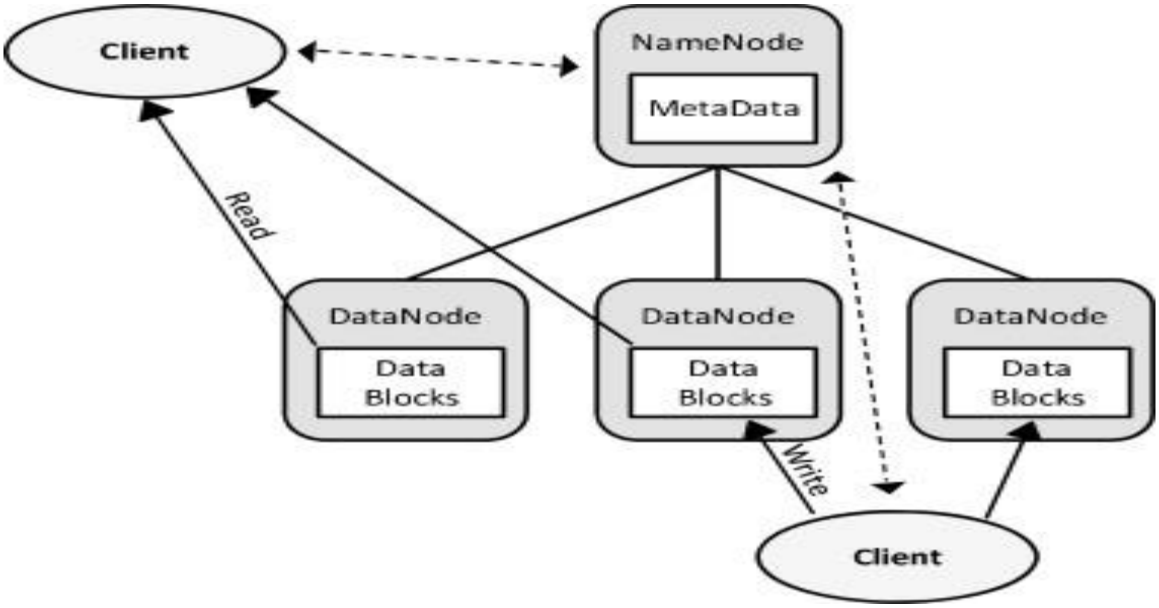


Figure 3. Hadoop File System and Data Communication Architecture [67].

Characteristics attributed to the Apache Hadoop software system are data integrity, scalability, and fault tolerance computing and processing. The Hadoop framework allows processing of large data sets across cluster of computers using simple programming model. It is designed to serve not only to single server but all thousands of machine, each offering local computation and storage. Further, several applications ranging from data warehousing to data flow oriented programming languages are implemented based on Hadoop [66, 68]. The Hadoop Cloud computing and Cloud Storage incudes new computing approaches based on Hadoop. The ability to perform parallel computations and other properties made Hadoop very popular.

Since initial release Hadoop changed constantly the version 2.2.0 on Oct 15, 2013 was developed. Besides bug fixes, new features and modules were developed and incorporated to process large amount of data.

3.2.2 APACHE FLUME

Apache Flume is a tool designed and implemented by Apache Software Foundation. The purpose of Flume software is to collect, aggregates, and move large log data files from many sources to a central storage in HDFS [69]. The use of Apache Flume is not limited to log data collection. It can also be used to transport large quantities of event data such as network traffic data, email messages and so on. The system requires: Java –Runtime Environment (Java 1.6 or later version), sufficient memory configuration, sufficient Disk Space and directory permission used by agent [69].

The basic architecture of streaming data is shown in Figure 4. It has three components, namely source, channel, and sink. The Figure illustrates how an external source like a web server sends events to flume source in a defined format. The event is stored in a channel or channels until it is used by flume sink. The sink removes the event and puts it in external storage like HDFS or forwards it to the flume source of next Flume Agent. The source and sink within the agent run asynchronously. Flume allows a user to build multi-hop-flows where events travel through multiple agents before reaching final destination. Flume provides end-to-end reliability of flow.

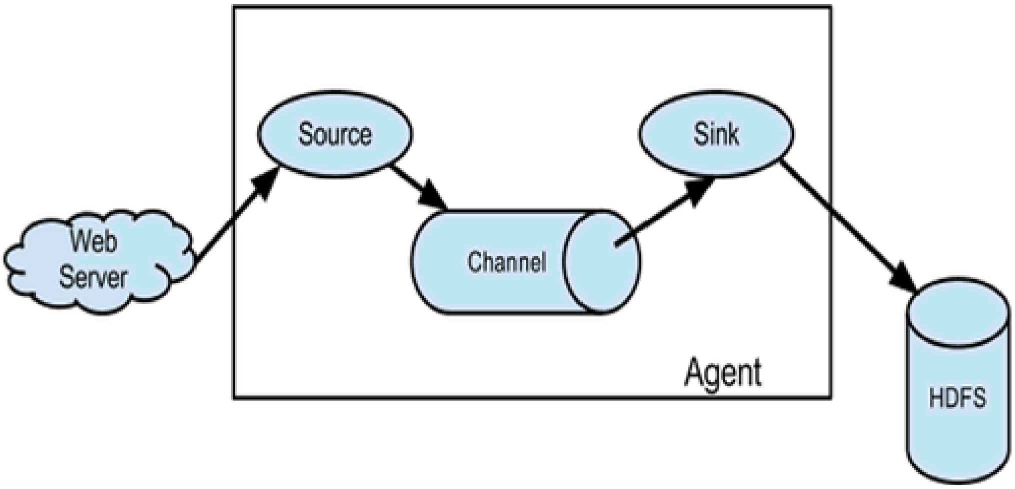


Figure 4. Basic Architecture on Streaming Data Flow [69].

In order to stream data on to HDFS from the data source, a flume agent should be created. A flume agent is a process that hosts the components through which events flow from an external source to next destination. For reliable source of events, transactional approach is used by flume. The events are staged in channel, which manages recovery. Flume supports a durable file channel, which is backed by the local file system, and also there is a memory channel, which simply stores the events in memory queue. Flume agent configuration has a text file that follows JAVA properties file format with hierarchical properties and settings. The configuration file includes properties of each source, sink and channel which are completed together to form data flow. By using a flow multiplexer, flume supports routing of an event to one or more destination. Output from Flume can be in JSON and Avro formats.

Flume can be used to retrieve tweets and store in the Hadoop file system. These files can be read by Hive. Figure 5 illustrates the dataflow from Flume to Hive. Twitter Streaming API provides access to Twitter’s global stream of data. It offers three types of streaming endpoints viz.

a. Public Streaming: Streams public data flowing through Twitter.

b. User Streaming: Single user’s twitter data.

c. Site Streaming: The multi-user version of user streams.

Flume is configured to capture the streaming data from one of the twitter end points. As seen before, flume stores data streamed from various sources into HDFS. The storage functional block of flume is called as HDFS Sink.

The data streamed from twitter is a continuous and large amount of data, which is channeled to HDFS Sink. To minimize the storage space this data is further compressed and then stored into HDFS. This compressed form of data is in the form of JSON Format (More on JSON Format is explained in the later section.).

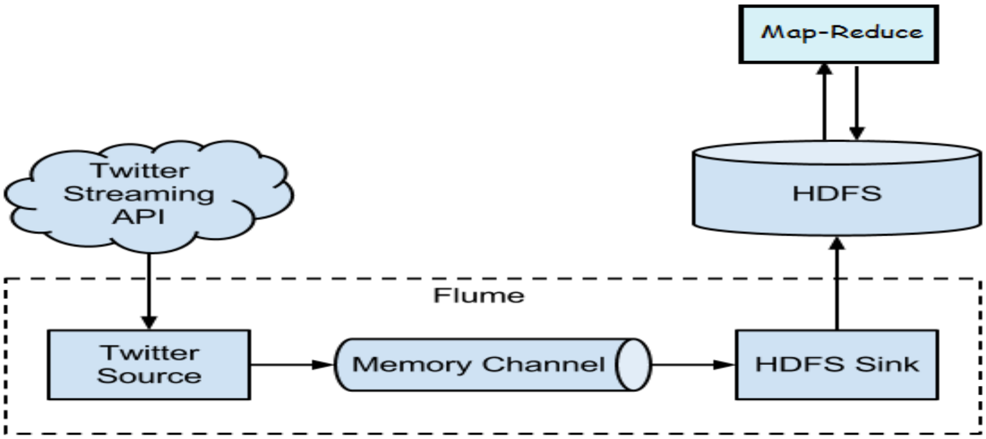


Figure 5. Dataflow diagram of Analysis of Twitter Stream Data [70].

3.4 HIDDEN MARKOV MODEL

The basic theory of Markov existed from the last 80 years but it is only recently that its application is used explicitly to solve processing problems. Markov models form a broad and flexible class of models with relatively easy analysis and straightforward interpretation. Hidden Markov Model (HMM) with discrete underlying state space and observations at discrete times has variable extensions. It is important for modeling time series data. They are used in almost all current speech recognition, in numerous applications in computation, in data compression and in other areas of artificial intelligence and pattern recognition. HMM is a powerful technique for solving problems, which has been augmented with three kinds of probabilistic information [40].

1. Each state has a probability with the machine starts.
2. Each transition, say from state p to q, has a probability that whenever the machine is in state p, it will go to state q.
3. Each output symbol c at each state q is labeled with the probability that the machine, if it is in state q, will output c.

   Formally, an HMM *M* is a quintuple (K, O, π, A , B) , where:

   • K is a finite set of states.

   • O is the output alphabet.

   • π is a vector that contains the initial probabilities of each of the states.

   • A is a matrix that represents the transition probabilities. A[p, q] = Pr(state q at time t | state p at time t - 1).

   • B, sometimes called the confusion matrix, represents the output probabilities.

   B [q,o] = Pr(output 0 | state q). Note that outputs are associated wi1h states (as in Moore machines).

In our research, HMM is used to model user interest in the Democratic and Republican parties based on the computed sentiments form the tweets. The model is used to predict the election as well as detect changes in the interest of Twitter users in the political parties.

3.3 JSON FORMAT

JSON or Java Script object Notation is way to store information in an organized easy manner. It is an open standard data interchangeable format. JSON was derived from JavaScript and ECMA script [71][72]. It is easily readable and writable for humans and machines. It is primarily used to transfer data over the network from server to client like XML. XML consumes lot of bandwidth over the network than a lightweight JSON. With JSON one can easily access the collection of data in a logical manner. JSON supports conventional data type that are similar to programmers of C family, Java, JavaScript, Python, etc. JSON's basic data types are:

- Number
- String
- Boolean
- Array and
- Object
- Null

Several programming languages provides APIs to parse and generate JSON data, given the above conventional data types.

JSON Schema defines the structure of the JSON data for validation, documentation and interaction control. JSON Schema is a contract to which JSON data should adhere similar to XSD for XML data. Below is the sample JSON Schema and its corresponding JSON data:

JSON Schema

```
{
  "$schema": "http://json-schema.org/draft-03/schema#",
  "name": "Product",
  "type": "object",
  "properties": {
            "id": {
```

```
        "type": "number",
            "description": "Product identifier",
            "required": true
        },
    "name": {
        "type": "string",
        "description": "Name of the product",
        "required": true
        }

    }
}
```

JSON Data:

```
        {
            “id” : 1,
            “name” : “Pencil”
        }
```

Loading data quickly and asynchronously is very important in today’s world and in this context JSON allows to overcome the cross - domain issue. Present Web browsers are working on native JSON encoding/decoding and it not only reduces security problem but also increase performance [73].

Below is a single JSON record, which is fetched from Twitter using flume:

```
{
        "filter_level":"medium",
        "retweeted":false,
        "in_reply_to_screen_name":null,
        "possibly_sensitive":false,
        "truncated":false,
        "lang":"en",
        "in_reply_to_status_id_str":null,
        "id":527594102289092608,
        "in_reply_to_user_id_str":null,
        "timestamp_ms":"1414623210871",
        "in_reply_to_status_id":null,
        "created_at":"Wed Oct 29 22:53:30 +0000 2014",
        "favorite_count":0,
        "place":null,

        "coordinates":null,

        "text":"MT\"@GlendaJazzey: Illegal Aliens Caught Stealing US Elections Via Felonies In New Studies http://t.co/LWLMC4Ib1W\"",

        "contributors":null,

        "geo":null,

        "entities":{

            "trends":[],

            "symbols":[],

            "urls":[{"expanded_url":"http://www.alipac.us/illegal-aliens-caught-stealing-us-elections-via-felonies-new-studies-3421/#.VFFviVwkOek.twitter",

                    "...":[90,112],

                    "display_url":"alipac.us/illegal-aliens\u2026",

                    "url":"http://t.co/LWLMC4Ib1W"}],

            "hashtags":[],

            "user_mentions":[{

                        "id":1895563525,

                        "name":"Glenda Slayton",

                        "indices":[3,16],
                        "screen_name":"GlendaJazzey",
                        "id_str":"1895563525"
                    }]
            },
        "source":"<a href=\"http://twitter.com/download/android\"
rel=\"nofollow\">Twitter for Android<\/a>",
        "favorited":false,
        "in_reply_to_user_id":null,
        "retweet_count":0,
        "id_str":"527594102289092608",
        "user":{
            "location":"",
            "default_profile":true,
            "profile_background_tile":false,
            "statuses_count":8976,
            "lang":"en-gb",
            "profile_link_color":"0084B4",
            "id":1597210508,
            "following":null,
"protected":false,
"favourites_count":3,
"profile_text_color":"333333",
"verified":false,
"description":"Conservative. Liberty & Freedom. Military/Vets Support. Pro-Israel. MT/RT are to inform.",
"contributors_enabled":false,
"profile_sidebar_border_color":"C0DEED",
"name":"Seth Salt",
"profile_background_color":"C0DEED",
"created_at":"Tue Jul 16 01:27:04 +0000 2013",
"default_profile_image":false,
"followers_count":5541,
"profile_image_url_https":"https://pbs.twimg.com/profile_images/527515265547112450/NPN0rq3W_normal.jpeg",
"geo_enabled":false,
```

"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png",

"profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png",

"follow_request_sent":null,

"url":null,

"utc_offset":-10800,

"time_zone":"Atlantic Time (Canada)",

"notifications":null,

"profile_use_background_image":true,

"friends_count":5541,

"profile_sidebar_fill_color":"DDEEF6",

"screen_name":"SethSalt",

"id_str":"1597210508",

"profile_image_url":"http://pbs.twimg.com/profile_images/527515265547112450/NPN0rq3W_normal.jpeg",

"listed_count":86,"is_translator":false

CHAPTER IV

FINDINGS

There are several types of elections. The most common one is presidential election, which selects one winner from several candidates. The other types of elections are general elections that select political parties and parliamentary election that selects representatives of parliament. It is harder to get better prediction in an election that has many candidates. Besides the number of candidates and the election types, the effect of other categories such as keywords selection and evaluation methods influence the prediction.

For Data collection, there are many methods on how to connect and collect tweets from twitter. The method of keyword search has been used in this thesis. The duration of data collection is one of the most important factors that could affect forecast accuracy. The duration of data collection was two weeks of pre-election and two weeks of post-election. Post-election data is used for sentiment change analysis only. Data collection periods were immediately preceding the election and three months after the election. The criterion used was user participation. We estimated that user activity will be higher close to the election. Keyword selection for data collection is another important factor as tweets collected depend on keywords used by researchers. The data collected from Twitter stream is classified into several groups based on several filters. The first goal in data filtering step is to reduce the noise (non-relevant data) in the dataset. Four steps of filtering have been conducted.

The data for computing the prediction results include several measures such as tweet count, distinct user count, retweet count and sentiment measure. The tweet count involves counting tweets both for pre-poll and post-poll in DEM, REP, DNR and NDR. Sentiment analysis is used to compute sentiment measures based peoples' reactions before and after the election. One of our objectives is to study the change in user interest in the two political parties after the election. Detailed description of the findings are given in the following subsections.

4.1 TWEET COUNT

In this the section, the findings of tweet based prediction has been explained. A total of 7655019 tweets were collected from the twitter data stream for pre-poll analysis, and 313956 tweets were found to be useful using first level of filtration. A total of 5500186 tweets were collected from the twitter data stream for post-poll analysis, but only 72621 tweets were found to be useful using the first level of filtration.

It was found that following the first filtration of pre-poll data, 109048 tweets belonged to REP, 62238 tweets belonged to DEM, 27027 tweets belonged to DNR and a large number (115643) belonged to NDR. Further, it was found that 30610 of the post-poll tweets belonged to REP, 11667 to DEM, 5151 to DNR and 25193 to NDR.

The second filtration was done by using negative keywords and phrases and the third filtration was done using the positive keywords and phrases for both events pre-poll and post-poll.

Figure 6 shows a comparison of the number of positive tweets in percentages in DEM, REP and DNR. It can be seen that REP has maximum percentage of positive tweets in comparison with DEM and DNR. There is a rise in the percentage of positive tweets from 29.58% to 31.38% in REP after the poll results in comparison with pre-poll tweets, but a slight fall of positive tweets is observed for DEM and DNR. It clearly indicates that people tweeted positively more for Republicans (REP) than for Democrats (DEM) and others (DNR).

Figure 7 shows negative tweets in percentage in DEM, REP and DNR. Negative tweets also were more for Republicans (REP) than for Democrats (DEM). This might be due to the fact that numbers of tweets in total were more in REP as people wanted to talk about Republicans (REP) more after poll results.

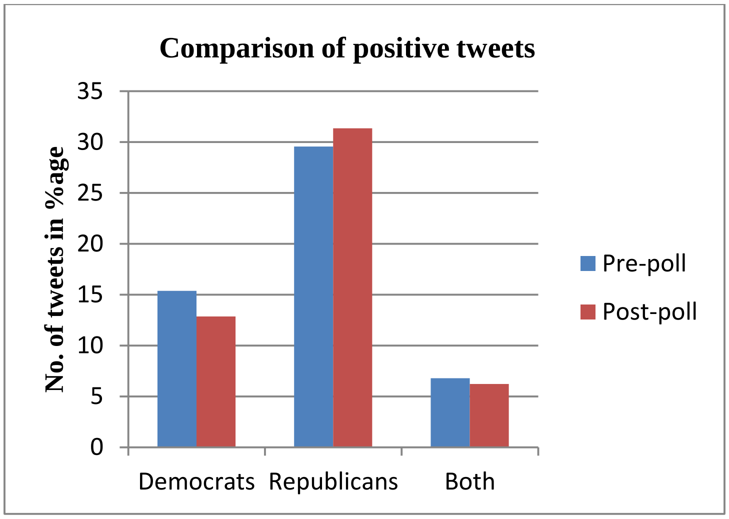


Figure 6. Comparison of positive tweets – DEM), REP and DNR

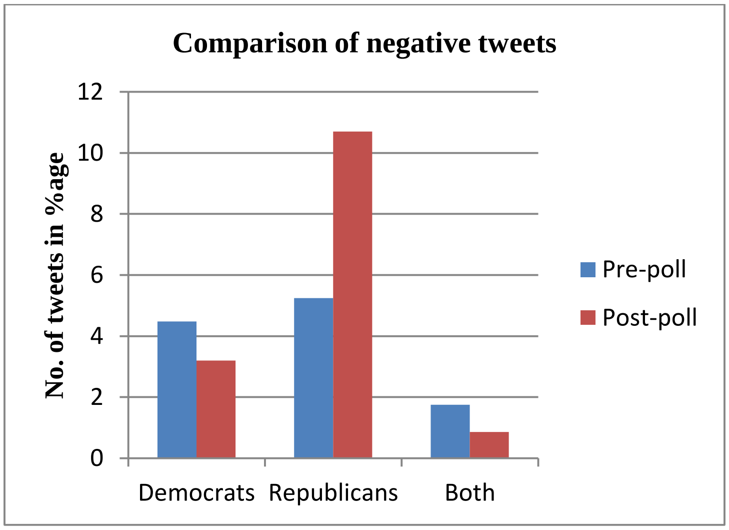


Figure 7. Comparison of negative tweets - DEM), REP and DNR.

A comparative analysis of Republicans and Democrats is performed using positive and negative pre-poll and post-poll data in REP, DEM, DNR and NDR.

The total positive tweets in REP, DEM, DNR and NDR are 92562, 48275, 21515 and 95521 respectively for pre-poll.

Figure 8 shows comparison of daily positive tweets for different parties during pre-poll. It can be seen that positive tweets of Republicans (REP) were more than Democrats (DEM). There was a significant number of positive tweets in REP compared to DEM and maximum number of positive tweets were found to be around 2nd November and 3rd November. This indicates that people were taking part in active politics and tweeted more about elections near the polling date that is the 4th November 2014. However we find relatively less tweets on 4th of November 2014.

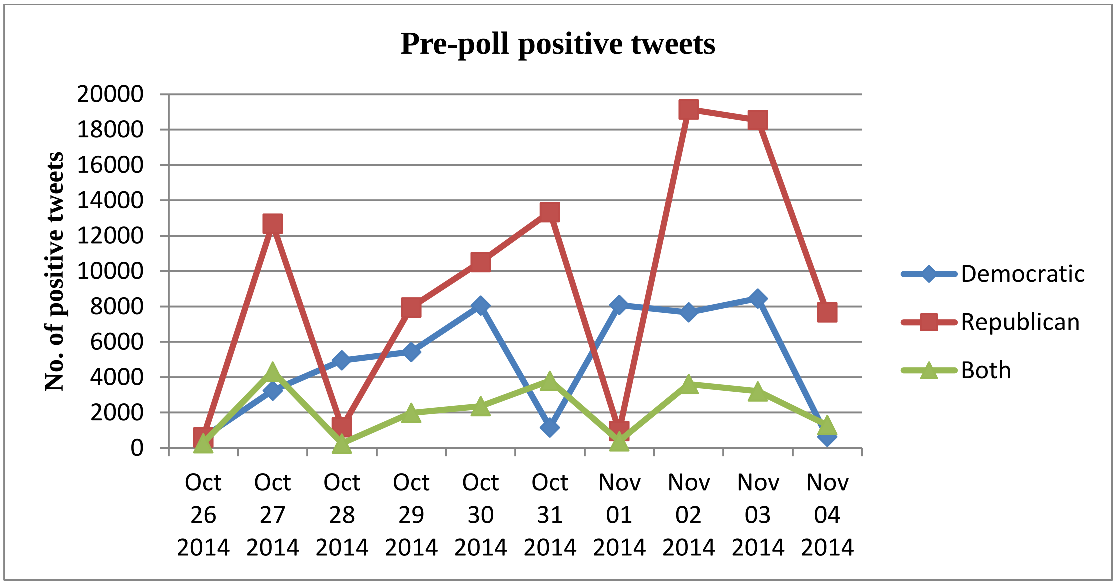


Figure 8. Pre-poll Positive tweets – DEM, REP and DNR.

The total negative tweets in REP, DEM, DNR and NDR are 16486, 13963, 5512 and 20122 respectively for pre-poll.

Figure 9 shows a comparison of negative tweets in REP, DEM and DNR. When REP and DEM are compared, a significant difference in the number of the negative Tweets was not seen. Clearly negative tweets were not able to give distinct prediction views as to whether Republicans (REP) or Democrats (DEM) would be preferred. The total positive tweets for post poll are 22791, 9339, 4523 and 21696 for REP, DEM, DNR and NDR.

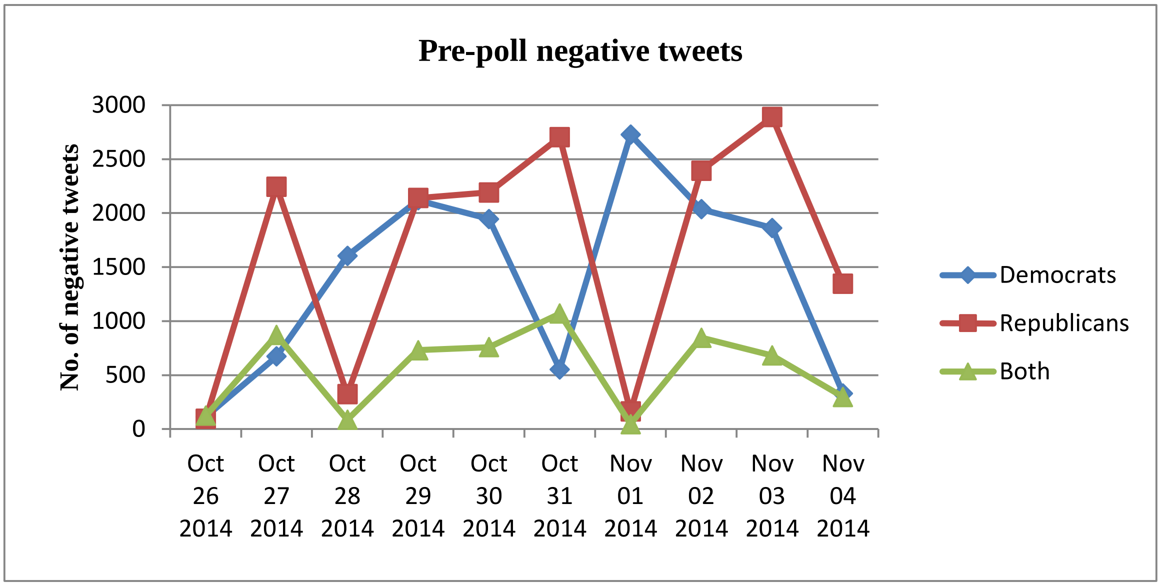


Figure 9. Pre-poll Negative tweets – DEM, REP and DNR.

Figure 10 shows a comparison of the number of post-poll positive tweets in REP, DEM and DNR at different days. Positive tweets are higher in REP than in DEM during the data collection period. A larger set of post-poll data would have given a better representation. Also if the data was collected just after the Election Day that is 4th November 2014, a more immediate reactions to the results would have been obtained.

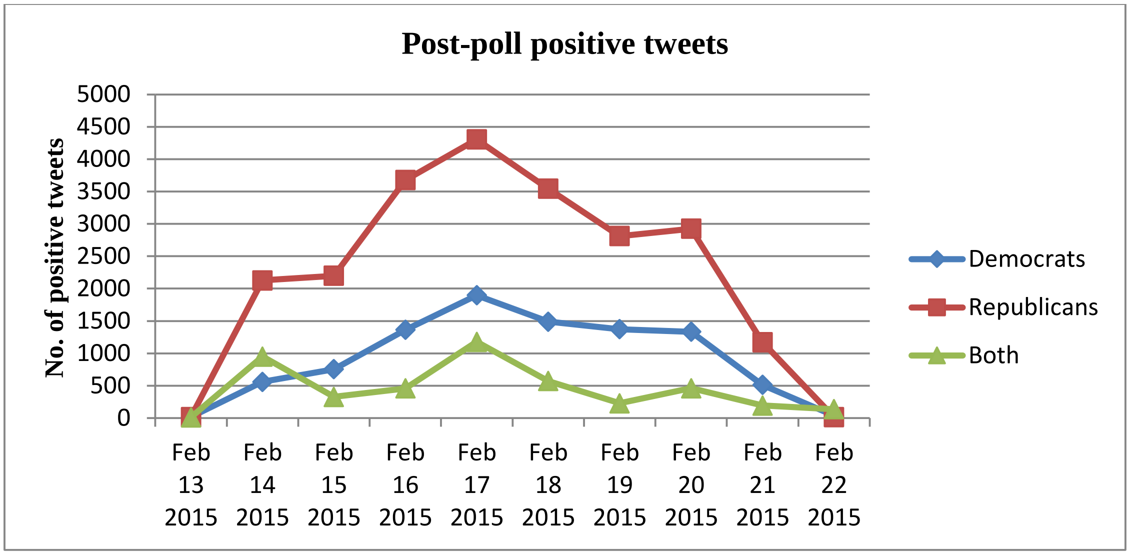


Figure 10. Post-poll Positive tweets – DEM, REP and DNR.

Figure 11 shows a comparison of the number of negative tweets in REP, DEM and DNR at different days. Negative tweets are significantly higher in REP than in DEM for few days. A larger post-poll data would have given a better representation. There is no obvious explanation for the spike in the negative tweets in REP. It may be due to some event which is not liked by the users regarding Republicans.

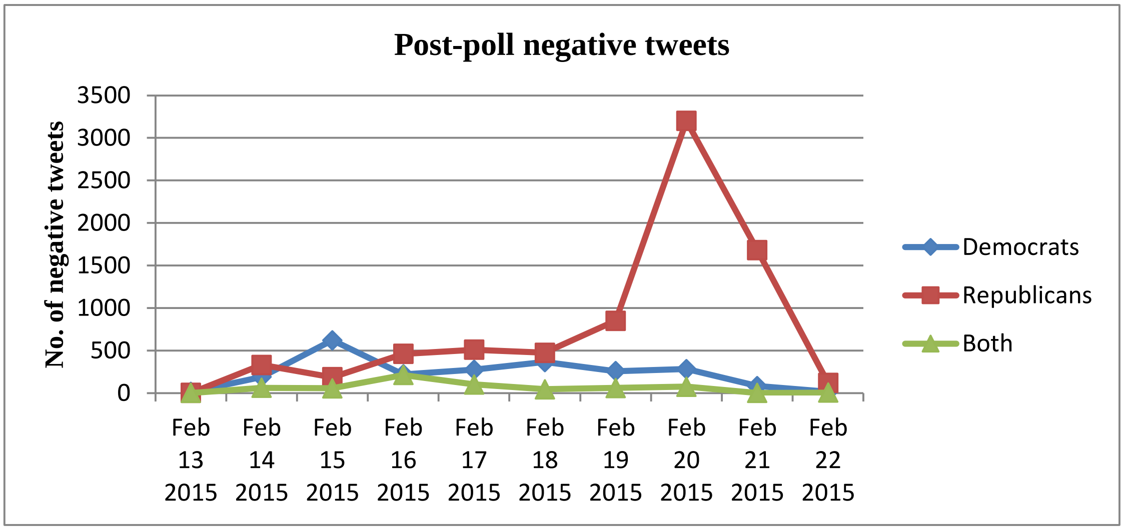


Figure 11. Post-poll Negative tweets – DEM, REP and DNR.

Figure 12 compares the number of positive and negative pre-poll tweets in NDR. It shows positive tweets more than negative tweets. Maximum positive tweets were found around 2nd November 2014 and 3rd November 2014. This indicates that the users who tweeted about candidates other than Republicans and Democrats were very positive about their candidates, especially immediately prior to the election. Figure 12 and 13 show that NDR included a large number of tweet data for both pre-poll and post-poll in case of positive tweets compared to negative tweets.

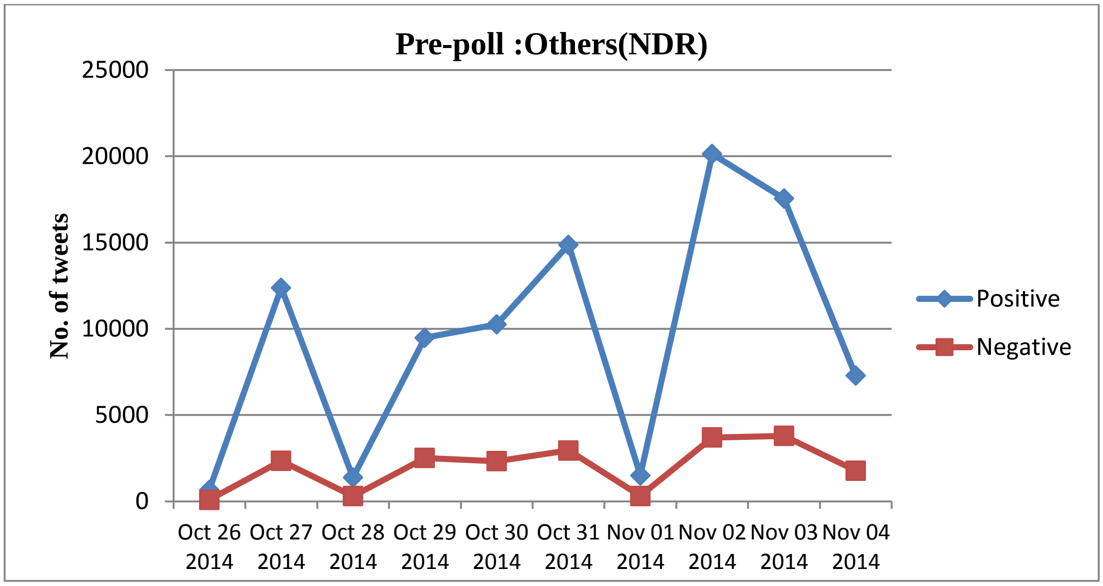


Figure 12. Pre-poll Positive and Negative tweets – NDR.

Figure 13 shows that post poll tweets in NDR giving positive tweets more than negative tweets. Tweet count analysis overwhelmingly favor republicans. As the Republican party won the election we conclude that tweet count is a reliable indicator for prediction.

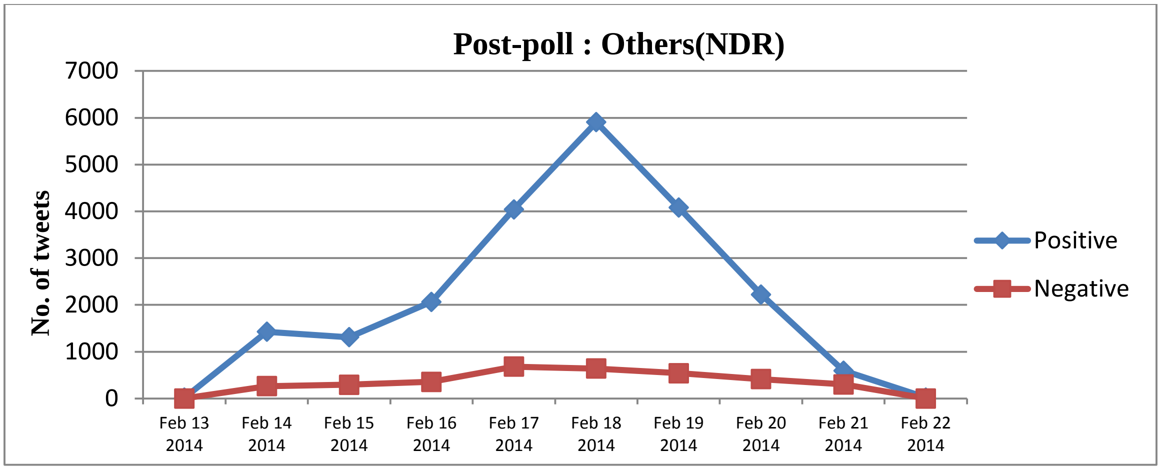


Figure 13. Post-poll Positive and Negative tweets – NDR.

4.2 RETWEETS

During pre-poll, total number of retweets was a large number. There were 54.02% retweets in comparison with useful data. Total positive retweets were 138021 in comparison with 31581 negative retweets. Also a comparison of retweets of positive tweets during pre-poll with post-poll retweets are given in Figure 14. Post-poll retweets are lesser in number than pre-poll retweets for all the three cases. There is a fall of 77.16%, 60.77% and 61.88% of number of retweets in REP, DEM and DNR respectively.

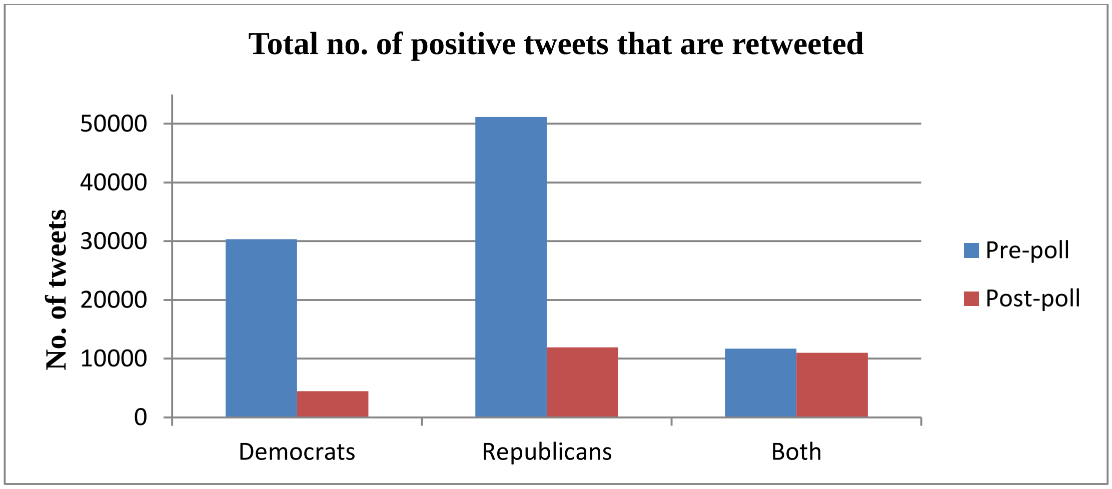


Figure 14. Total number of positive tweets that are retweeted.

Figure 15 shows total number of negative tweets that are retweeted. Negative tweets that are retweeted are more in REP than in DEM and DNR. Interestingly negative tweets were lesser in the post -poll data for all three categories than pre-poll data. Negative retweets was reduced by 44.73% for REP, 82.12% for DEMs and 92.25% for DNR in comparison to pre-poll. This indicated that people started retweeting less in the post-poll for both negatives and positive retweets. It is not obvious why the decrease in negative retweets of REP is much smaller than that of the DEM. Further analysis may provide indicators for the phenomenon. Pre-poll retweet data favors a Republican win.

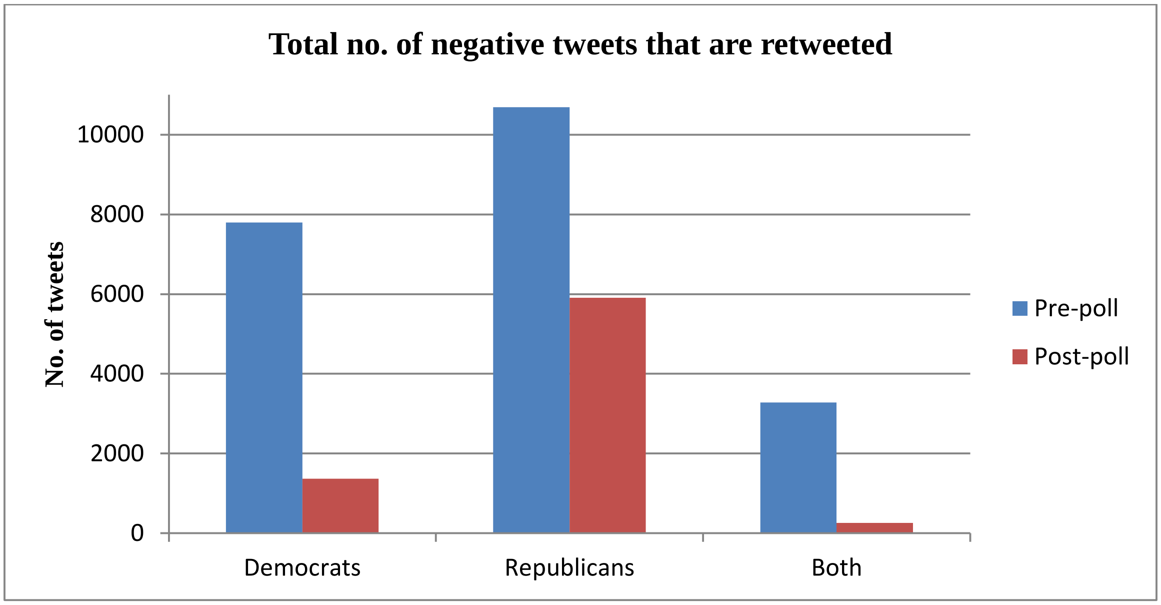


Figure 15. Total number of negative tweets that are retweeted.

4.2.1 COMPARISON OF ORGINAL TWEETS AND RETWEETS FOR PRE-POLL AND POST-POLL

As shown in the figures below, the number of tweets is higher than the number of retweets. This shows that the information diffusion rate is low. Tweet counts are low for all categories some days. We are unable to ascertain any specific reason for the low tweets. Figure 16(a) compares positive tweets and retweets during pre-poll in REP. Maximum number of tweets/ retweets was found around 2nd November 2014 and 3rd November 2014. This occurs in all categories. This may be due to voters' interest in politics as polling day neared. It can be noted that we have large number of retweets in comparison to useful tweets, even though less than the original tweets. Retweet pattern is similar to the tweet pattern. Based on this analysis, we conclude that retweets are a very important parameter for prediction analysis of elections. Retweet count seem to favor Republicans.

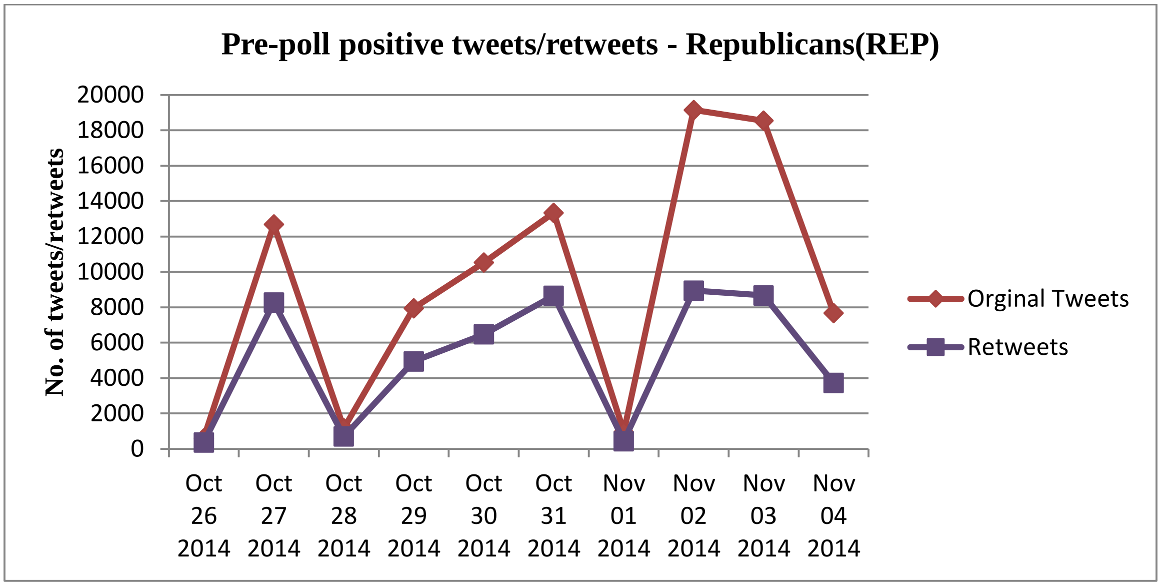


Figure 16(a). Pre-poll Positive Tweets – Original Tweets and Retweets for REP.

Figure 16(b) shows comparison of negative tweets and retweets during pre-poll in REP. The negative tweets and retweets do show similar pattern. There is increase in the number of tweets and retweets around 2nd November 2014 and 3rd November 2014. Again, it can be seen that retweets, which are negative in nature, form a large portion of tweet data for negative tweets.

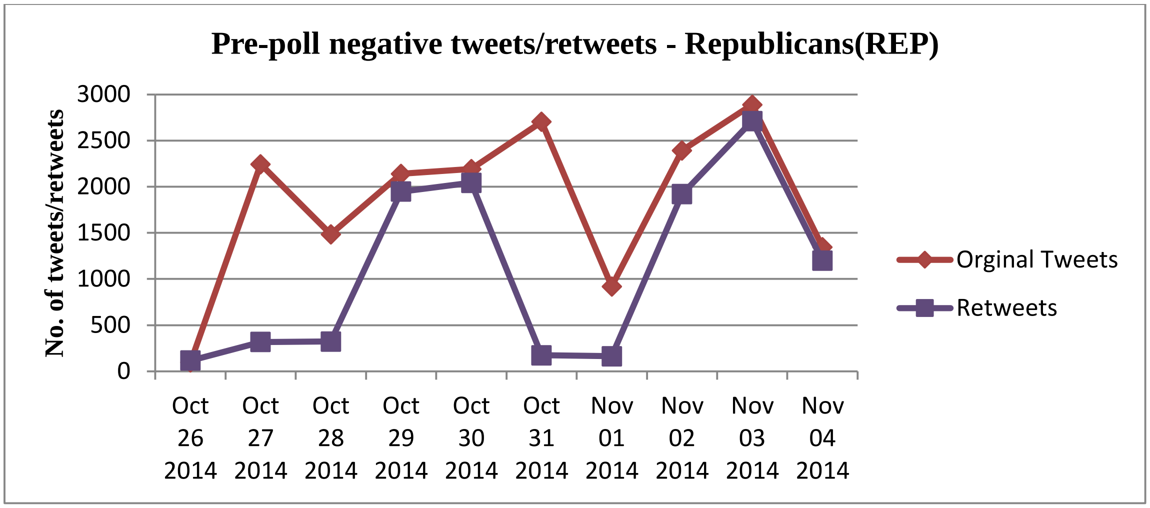


Figure 16(b). Pre-poll Negative Tweets – Original Tweets and Retweets for REP.

Figures 17(a) and 17(b) show positive and negative tweets and retweets during pre-poll period for DEM. Maximum positive tweets and retweets occur around 1st November 2014 and 3rd November 2014, but for negative tweets and retweets, it differs. Based on Figures 16 and 17, retweets for the REP category seems to be higher.

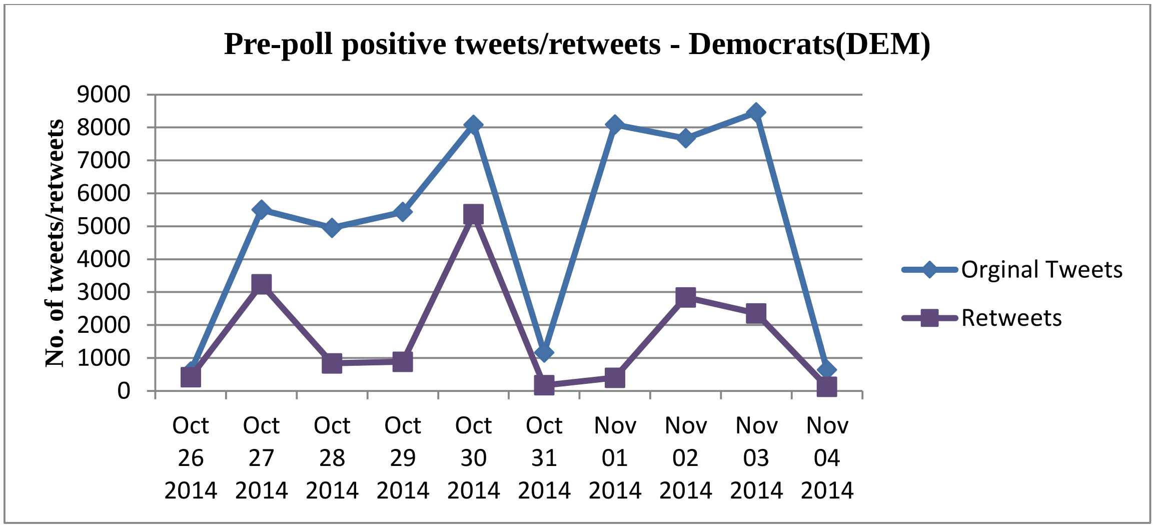


Figure 17(a). Pre-poll Positive Tweets – Original Tweets and Retweets for DEM.

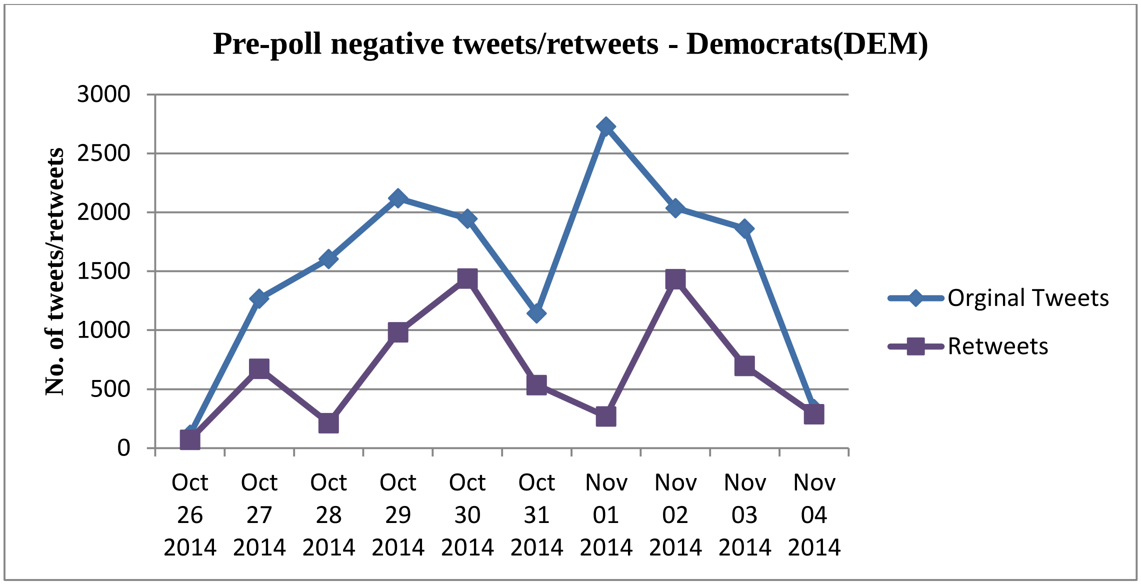


Figure: 17(b). Pre-poll Negative Tweets – Original Tweets and Retweets for DEM.

Figures 18(a) and 18(b) show pre-poll tweets and retweets in DNR (Tweets referring to both Republicans and Democrats) of positive and negative tweets/retweets respectively. Retweets seem to track the original tweets in count, but lower.

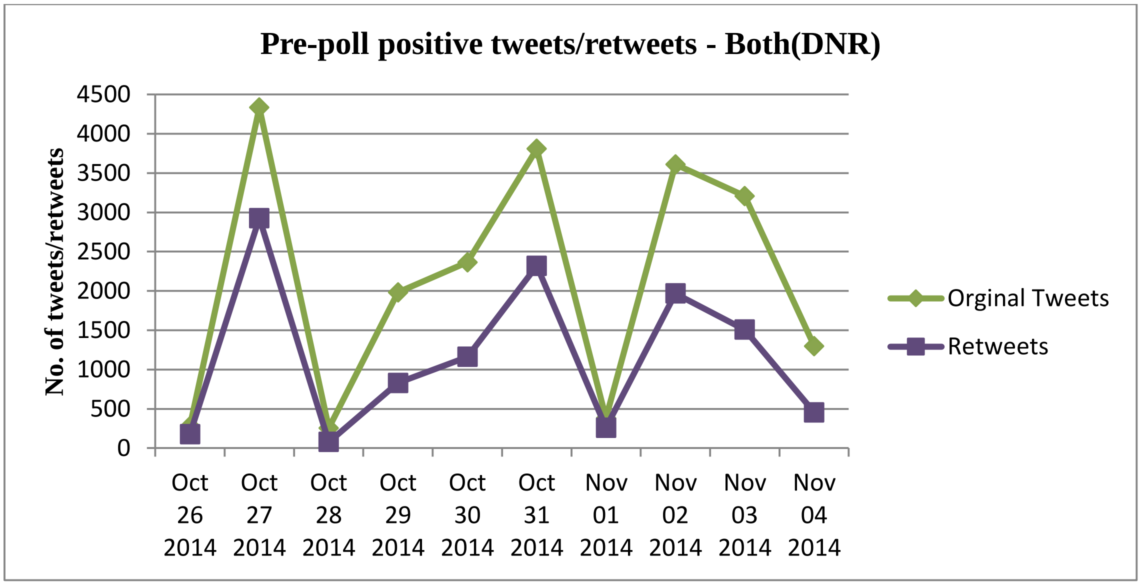


Figure 18(a). Pre-poll Positive Tweets – Original Tweets and Retweets for DNR.

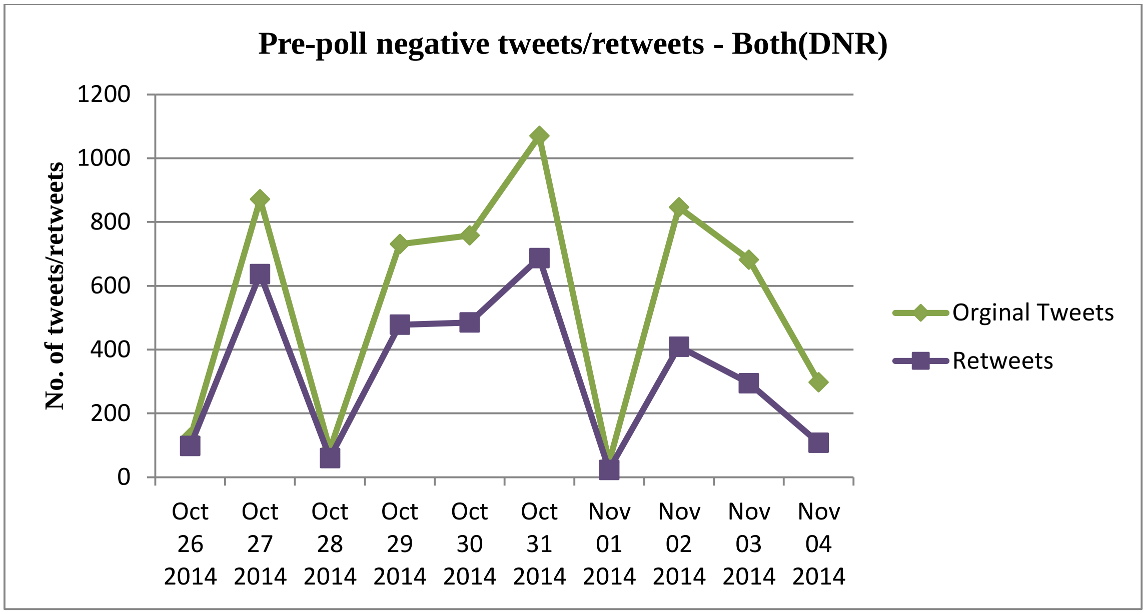


Figure 18(b). Pre-poll Negative Tweets – Original Tweets and Retweets for DNR.

Figures 19(a) and 19(b) show respectively the number of positive and negative tweets and retweets for the NDR category during the pre-poll period. It can be observed that the number of positive and negative tweets are high on 2nd November 2014 and 3rd November 2014.

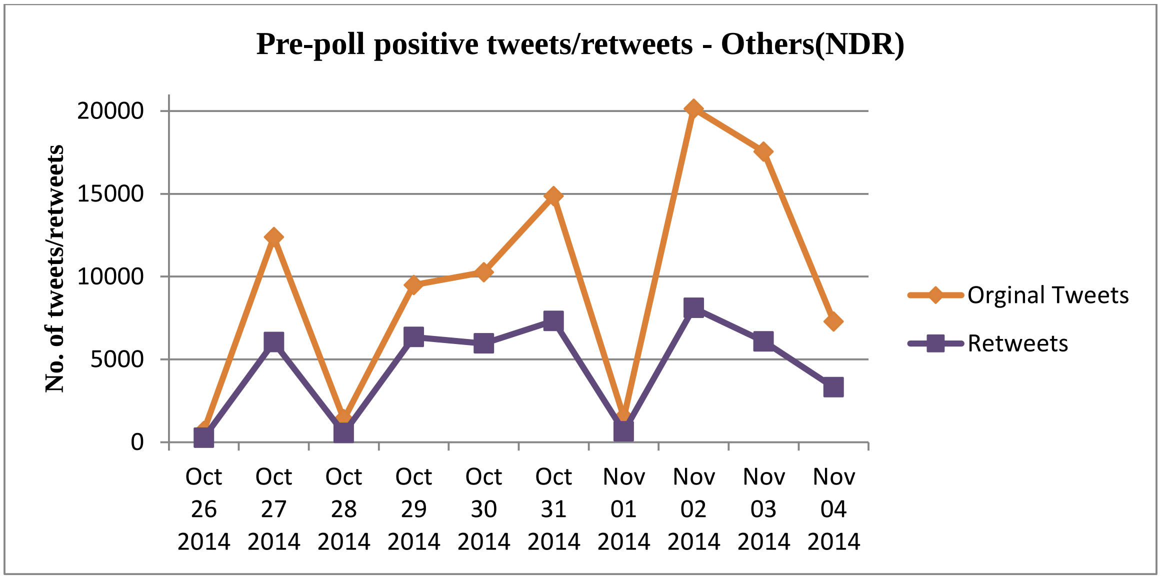


Figure 19(a). Pre-poll Positive Tweets – Original Tweets and Retweets for NDR.

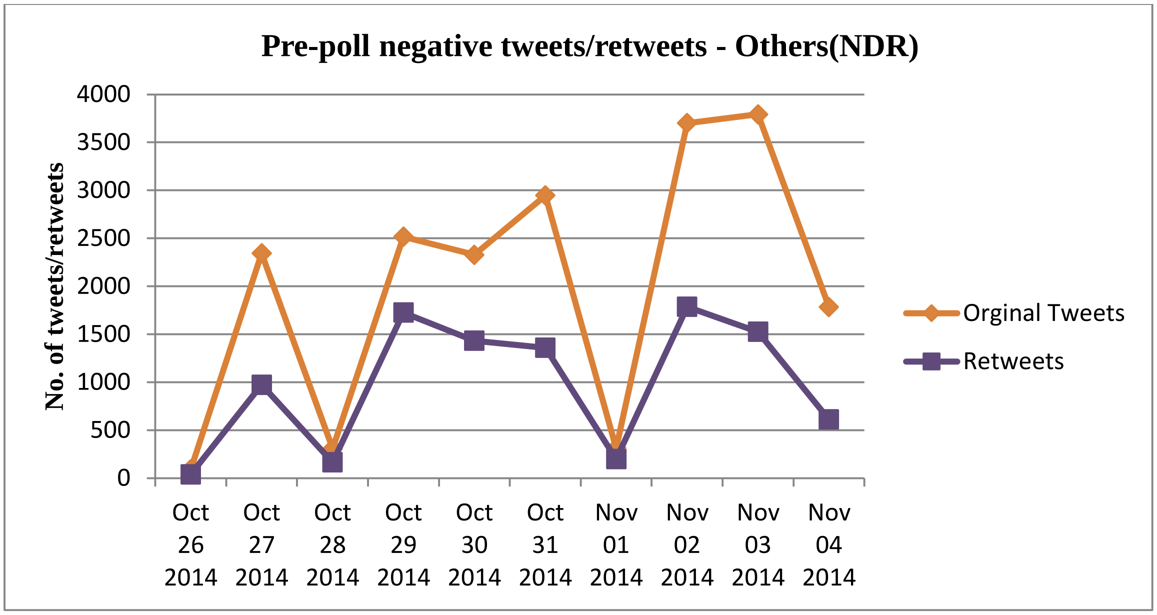


Figure 19(b). Pre-poll Negative Tweets – Original Tweets and Retweets NDR.

Figures 20(a) and 20(b) show the number of positive and negative tweet/retweets for post-poll for REP. We are unable to explain the spike in negative tweets and retweets. Negative retweet count is close to the original tweet count.

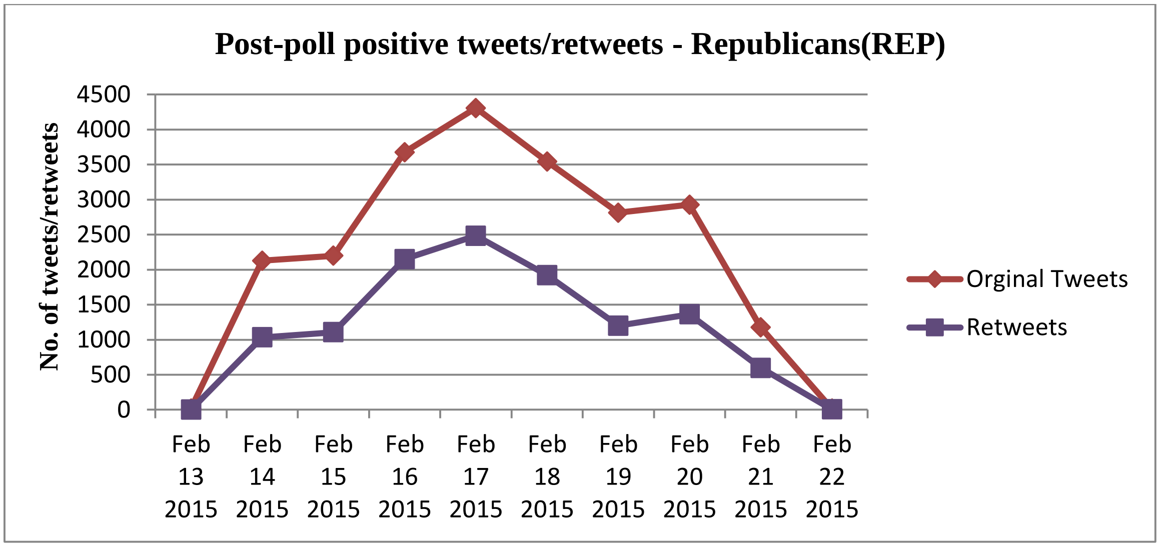


Figure 20 (a). Post-poll Positive Tweets – Original Tweets and Retweets for REP.

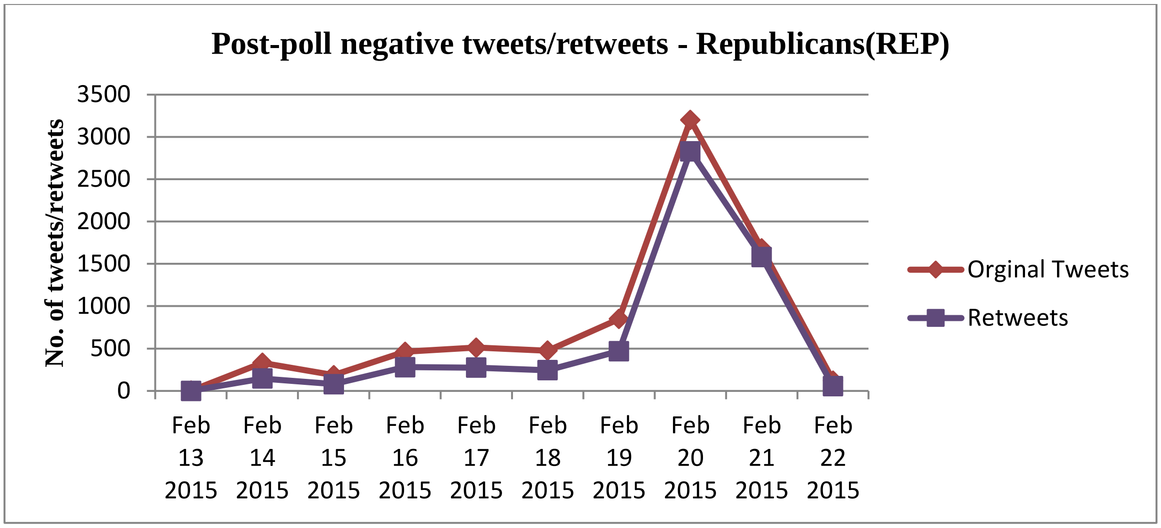


Figure 20 (b). Post-poll Negative Tweets – Original Tweets and Retweets for REP.

Figures 21(a) and 21(b) show the number of positive and negative tweets and retweets during the post-poll period for DEM. There seems to be more tweets some days than the other. Just as in the case of the Republican negative tweets, the reasons are not evident for the behavior. Data for a longer period might explain the behavior.

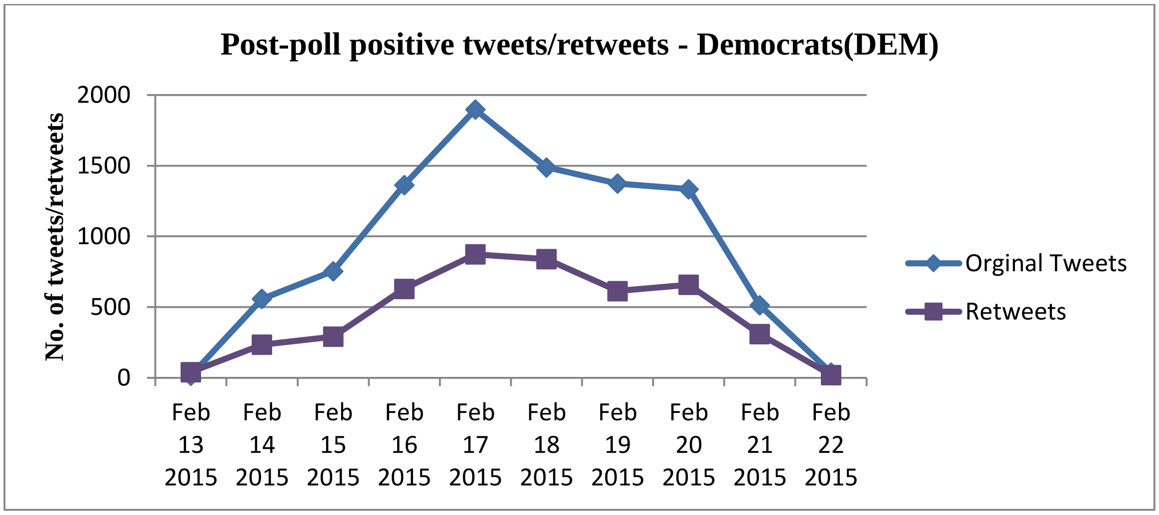


Figure 21(a). Post-poll Positive Tweets – Original Tweets and Retweets for DEM.

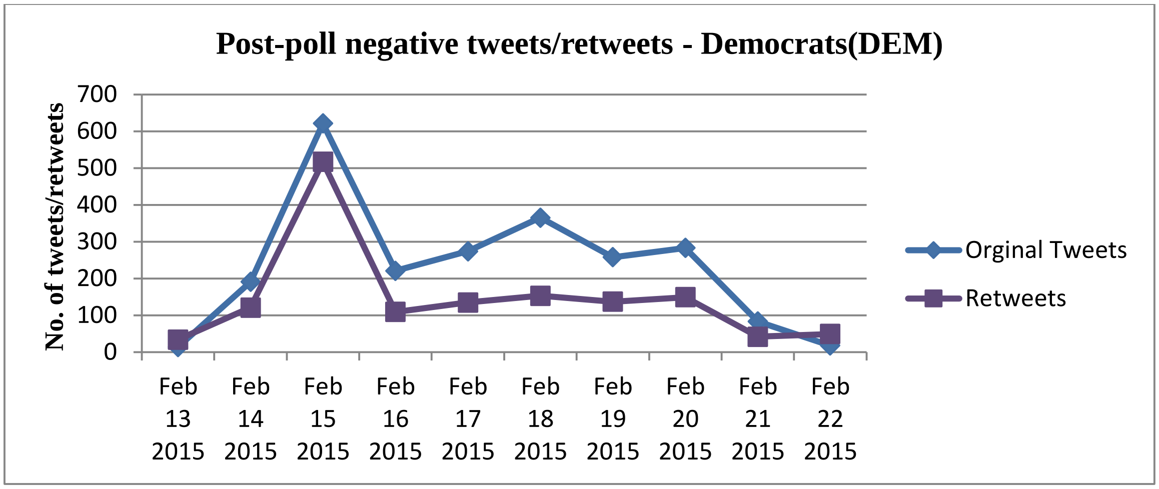


Figure 21(b). Post-poll Negative Tweets – Original Tweets and Retweets for DEM.

The DNR category contains all the tweets that refer to both Democrats and Republicans. However, the contribution of each in this category cannot be determined accurately. Further research is needed to obtain better insight of each party contribution towards positive tweets and negative tweets in pre-poll and post-poll analysis.

Figures 22(a) and 22(b) show the number of positive and negative tweets and retweets for post-poll for DNR category. Retweet count seems to track tweet count and both are decreasing.

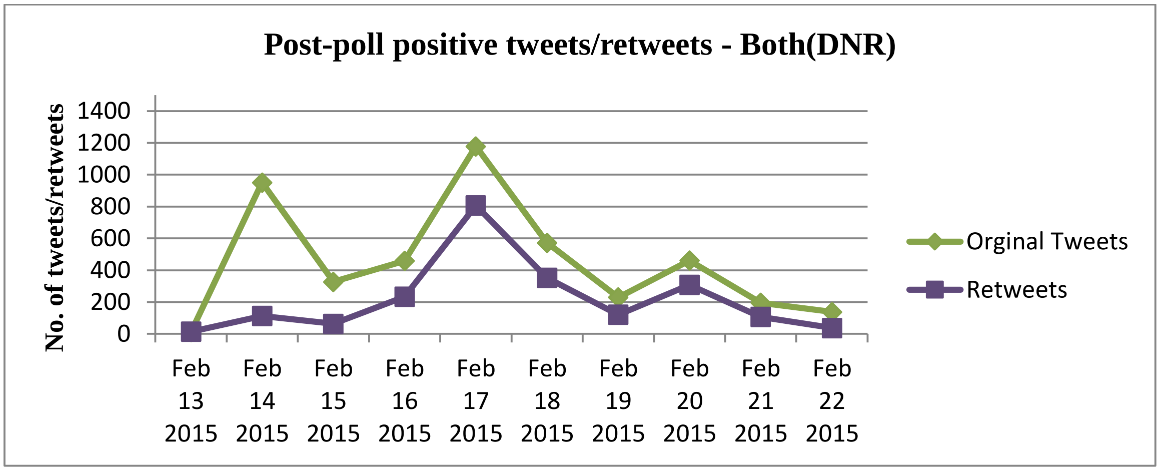


Figure 22(a). Post-poll Positive Tweets – Original Tweets and Retweets for DNR.

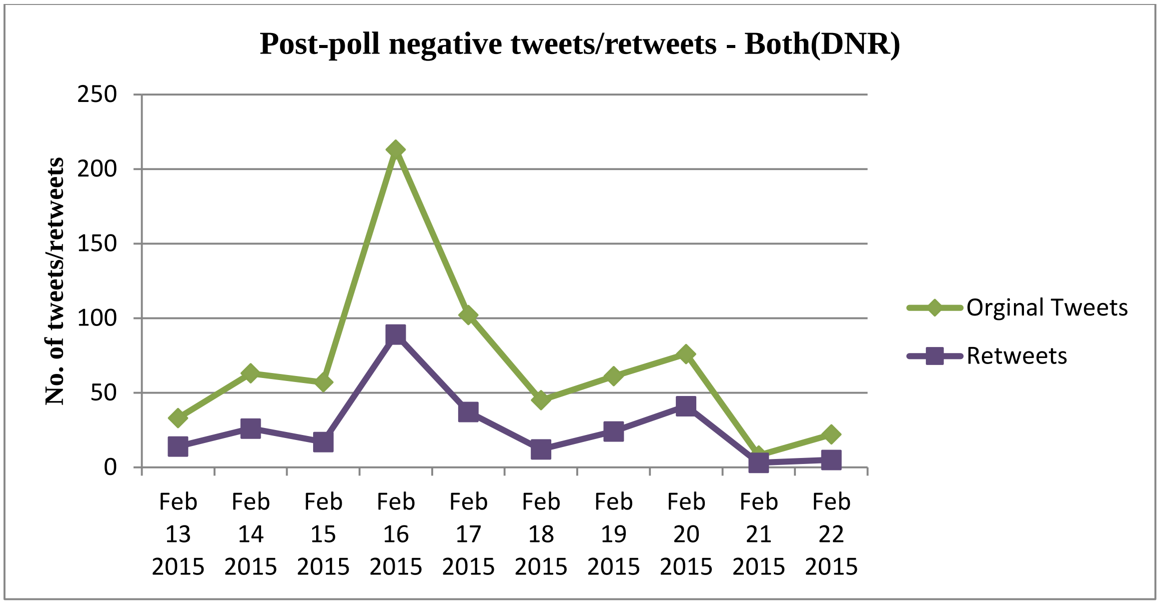


Figure 22(b). Post-poll Negative Tweets – Original Tweets and Retweets for DNR.

The number of original tweets for both positive and negative tweets are more than the number of retweets in case of NDR shown in Figure 23(a) and Figure23(b).

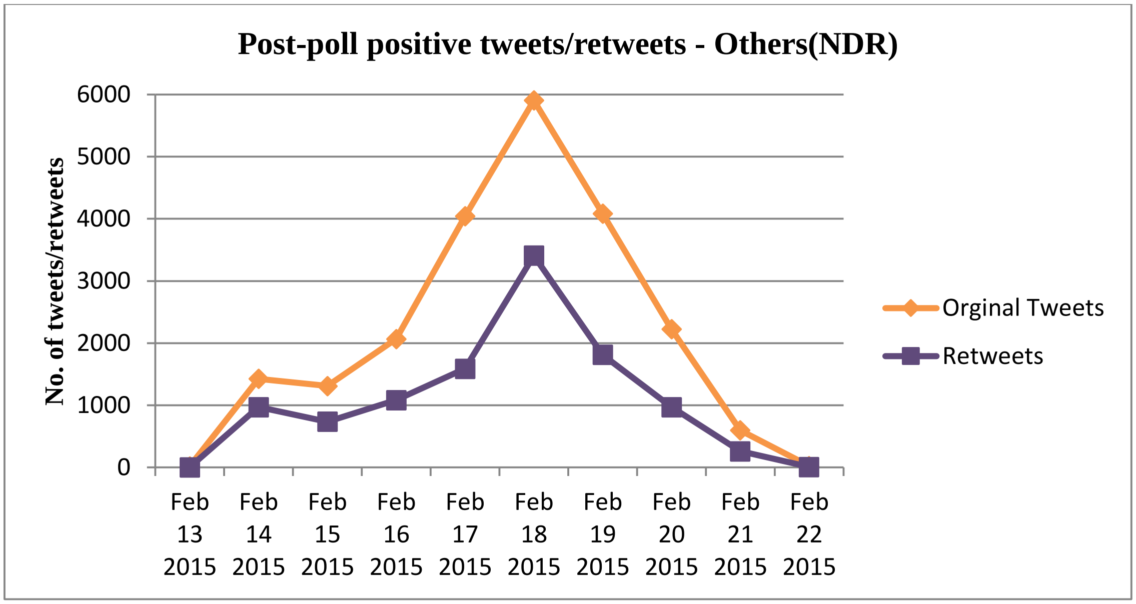


Figure 23(a). Post-poll Positive Tweets – Original Tweets and Retweets for NDR.

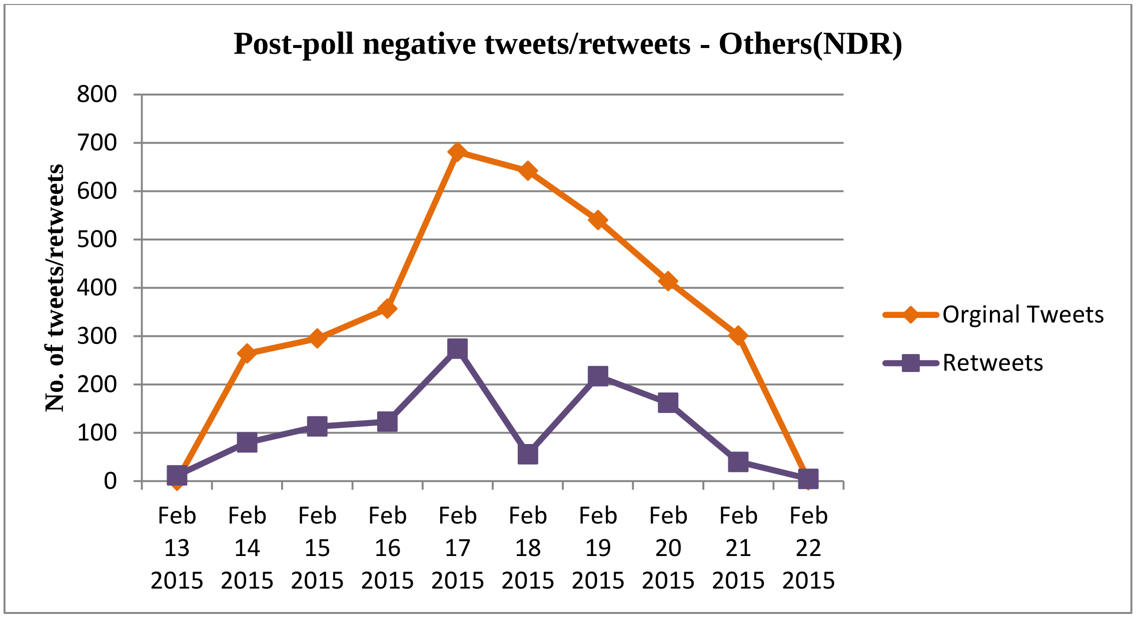


Figure 23(b). Post-poll Negative Tweets – Original Tweets and Retweets for NDR.

4.2.2 RETWEET FREQUENCY

In this section, we analyze the characteristic of the retweets. Related information are shown in Figures 24-36. Figure 36 provides a comparison of tweet count and retweet counts for both pre-poll and post-poll data collection periods. It should be observed that tweeting activity is significantly lower after the election. Figures 24, 26, 27, 29, 30, 32, 33 and 35 show frequency of top ten and bottom ten retweets for the four categories DEM (tweets containing only Democrats), REP (tweets containing only Republicans), DNR (tweets containing both Democrats and Republicans), and NDR (tweets containing neither Democrats nor Republicans). The x-axis in all cases represent the tweet and the y axis represent the frequency of retweet during the data collection period. It can be observed that frequency shows a linear downward trend for all categories. Figures 25, 28, 31 and 34 compare the strength of retweets corresponding to DEM and REP during pre-poll and post-poll data collection periods. Both positive and negative retweets are higher in REP which indicate more people are interested in the Republican party.

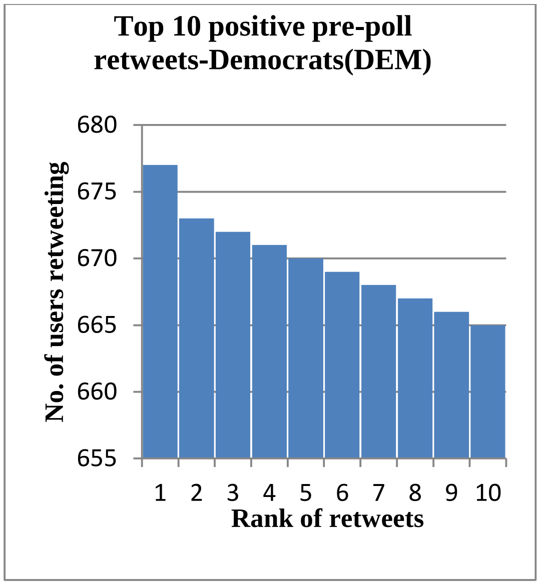


Figure 24(a). Top 10 positive pre-poll retweets for DEM.

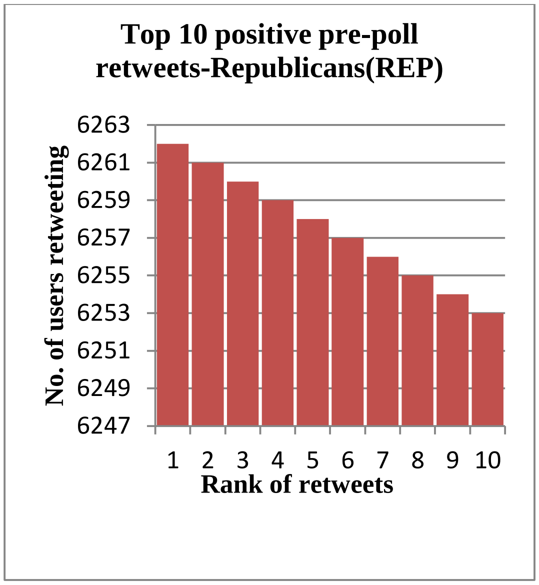


Figure 24(b). Top 10 positive pre-poll retweets for REP.

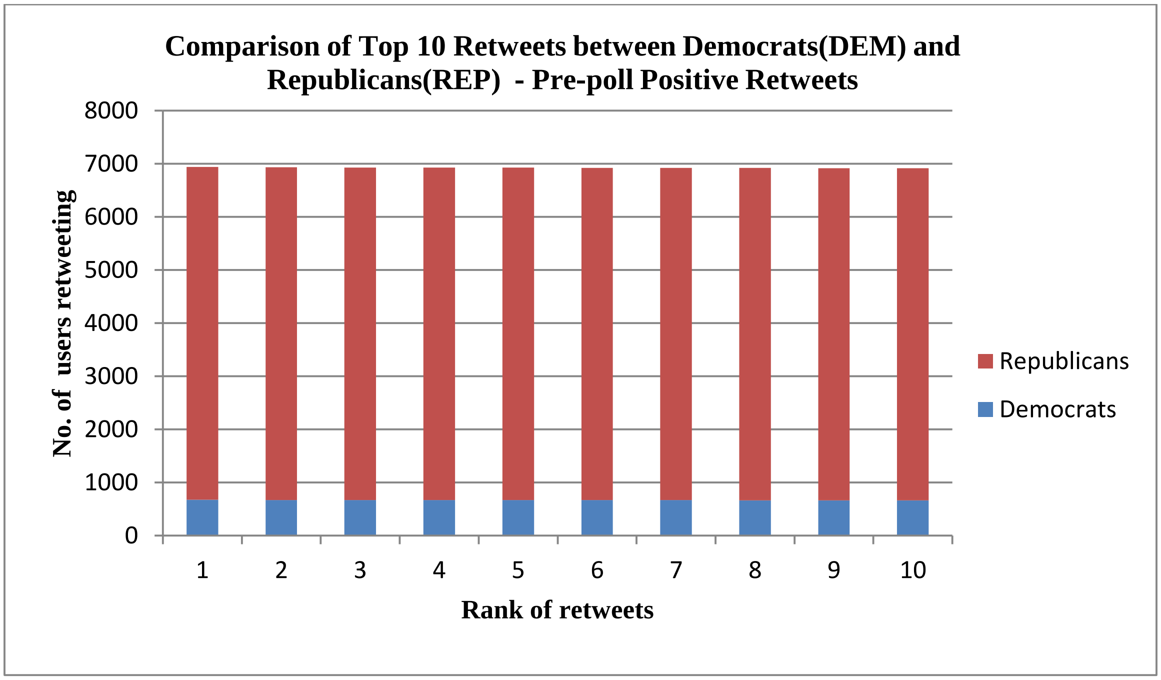
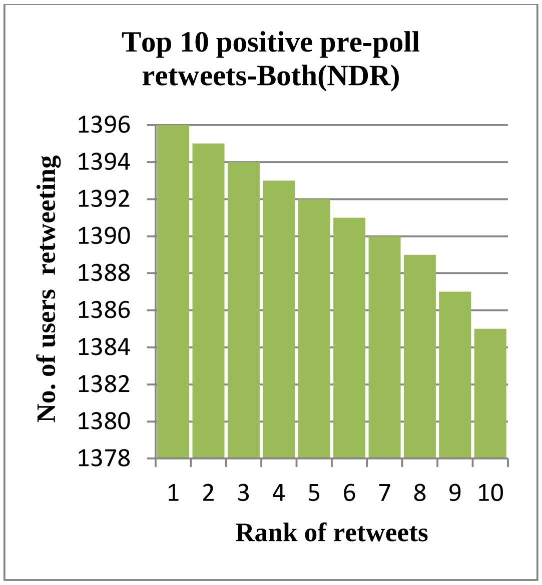


Figure 25. Top 10 positive retweets: DEM and REP – Pre-poll.



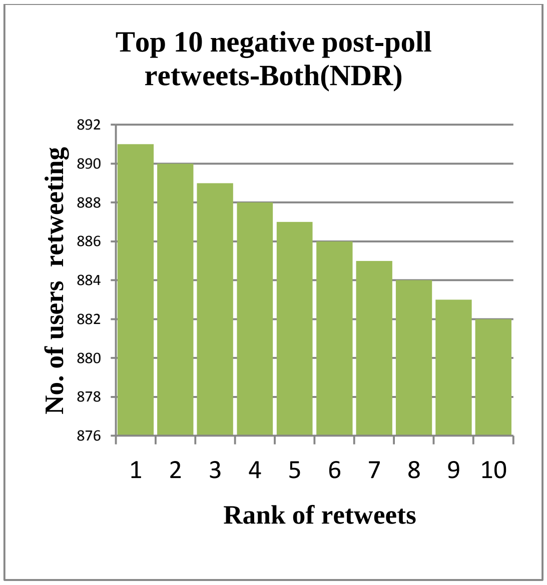
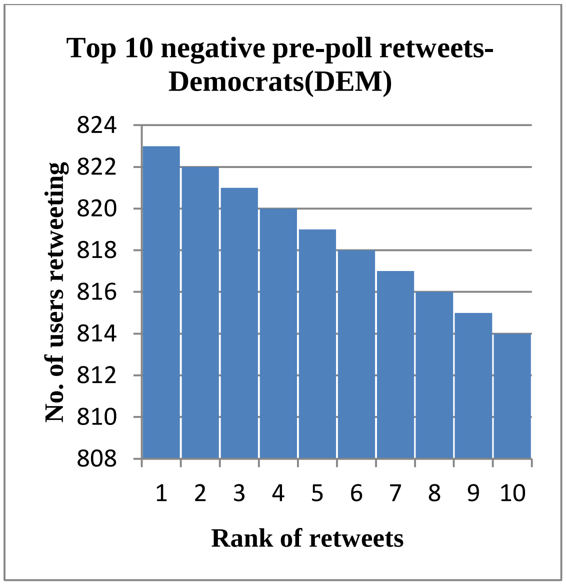


Figure 26(a). Top 10 positive pre-poll retweets for DNR.

Figure 26(b). Top 10 negative pre-poll retweets for DNR.



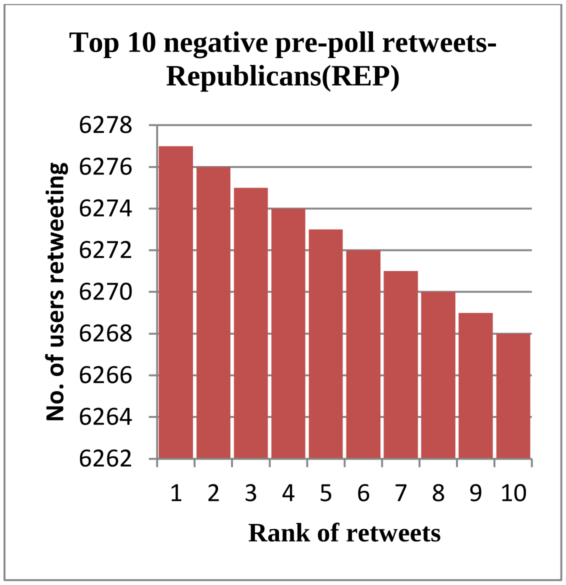


Figure 27(a). Top 10 negative retweets for DEM.

Figure 27(b). Top 10 negative retweets for REP.

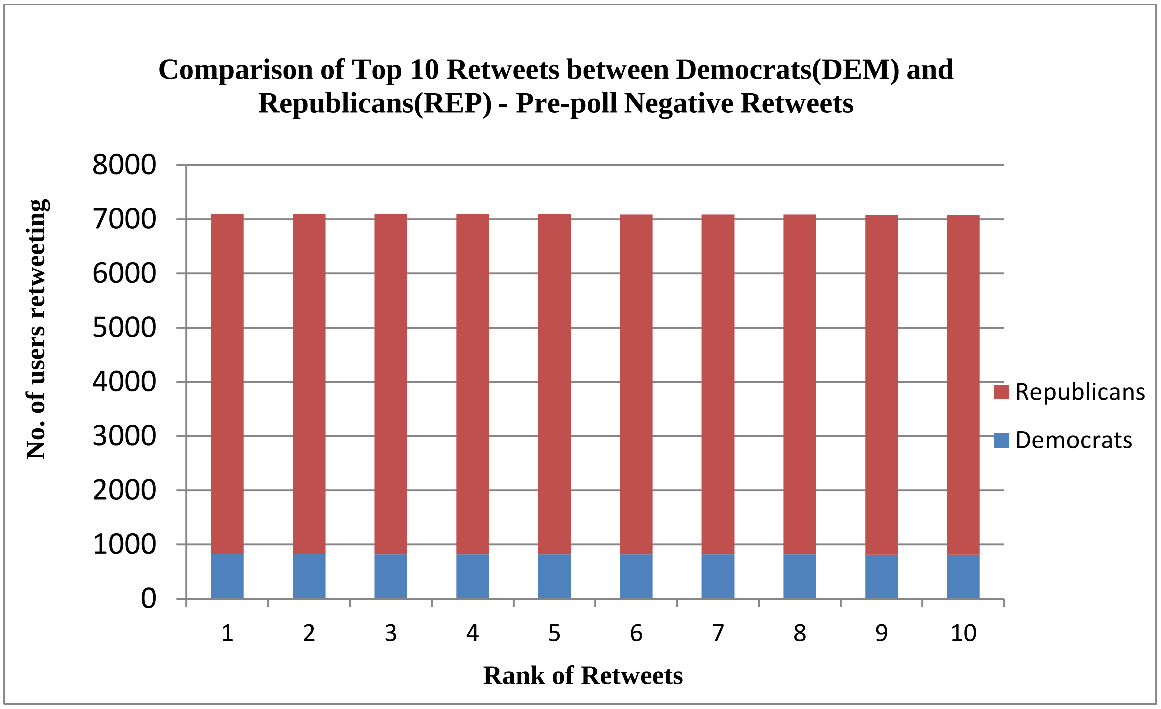
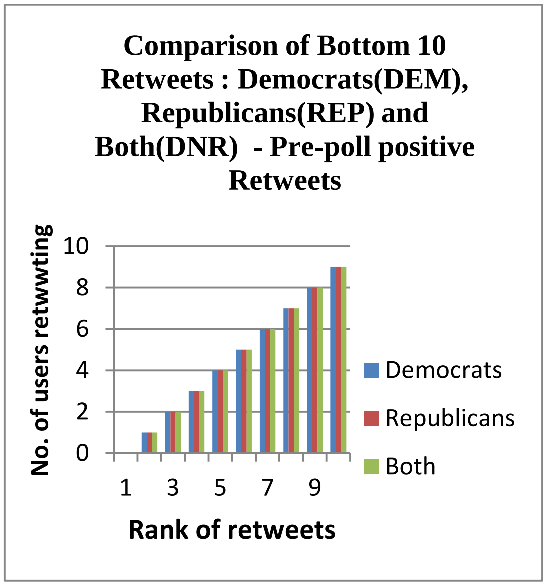


Figure 28. Top 10 negative retweets between DEM and REP – Pre-poll.



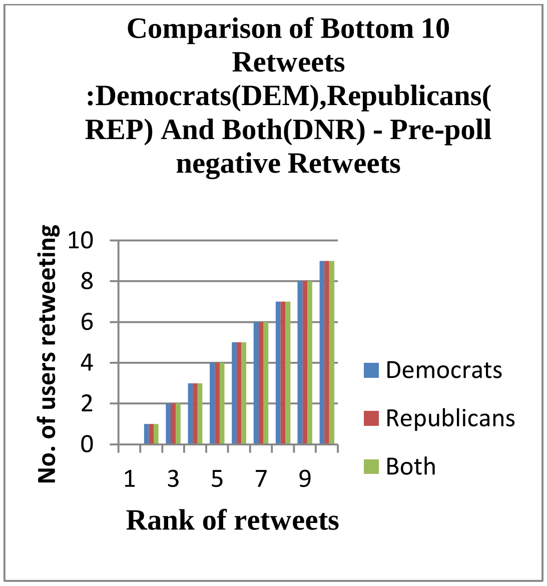


Figure 29(a). Bottom 10 positive pre-poll retweets: DEM, REP and DNR.

Figure 29(b). Bottom 10 negative pre-poll retweets: DEM, REP and DNR.

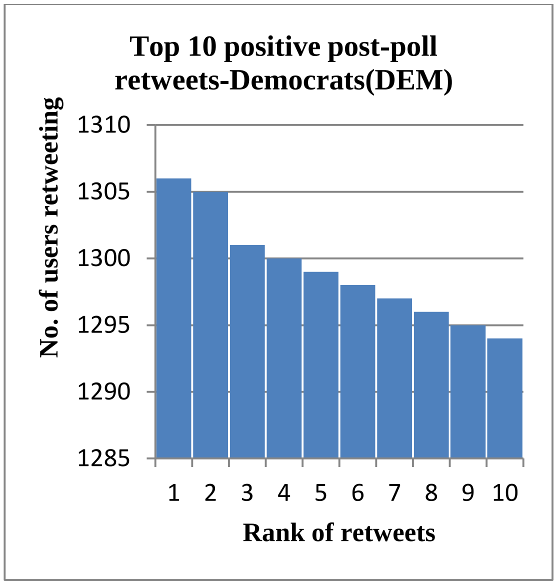


Figure 30(a). Top 10 positive post-poll retweets for DEM.

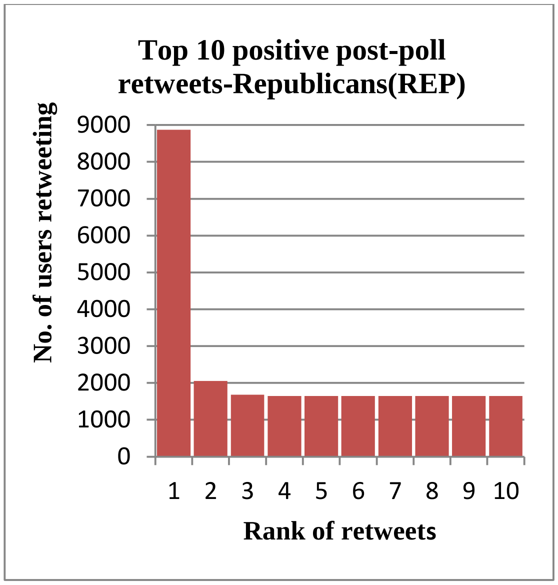


Figure 30(b). Top 10 positive post-poll retweets for REP.

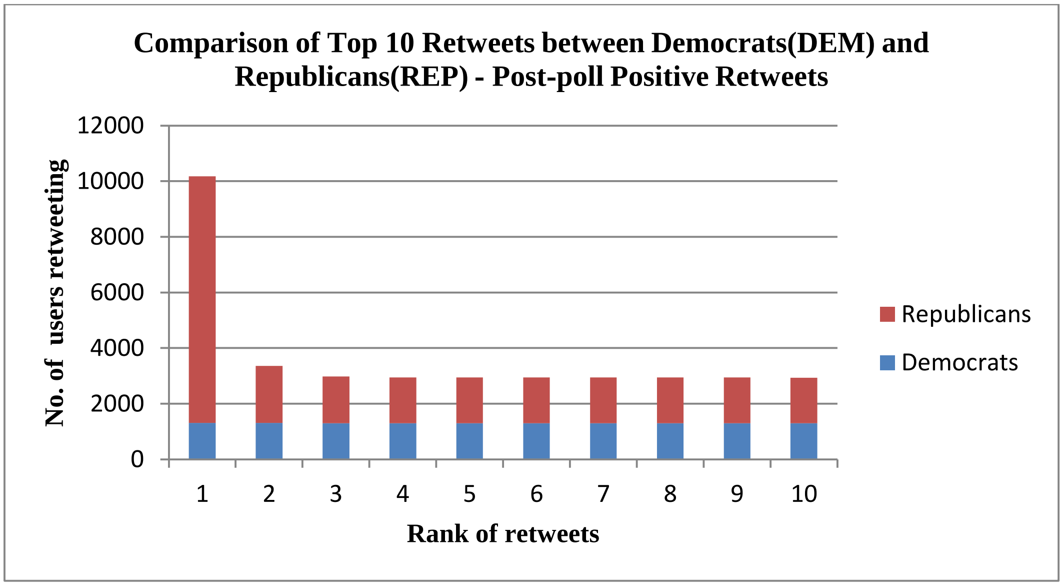
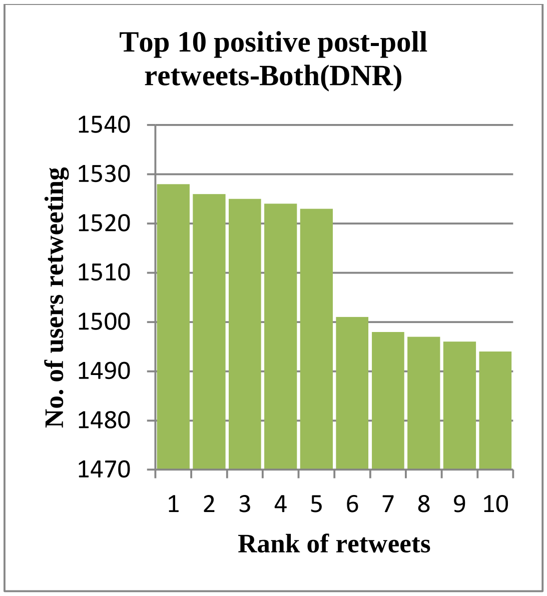


Figure 31. Top 10 positive retweets: DEM and REP – Post-poll.



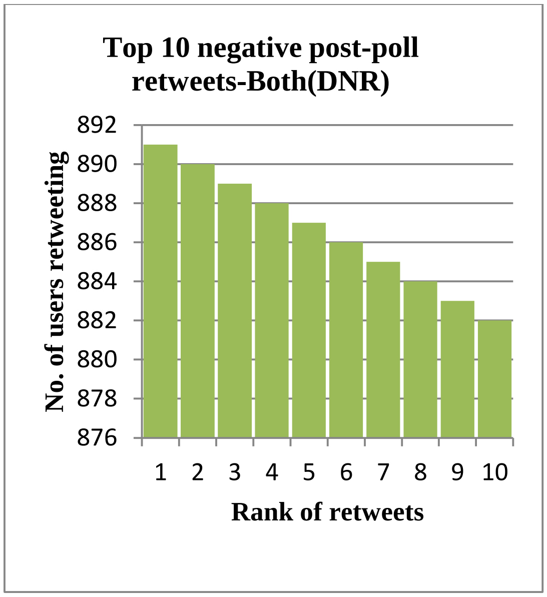
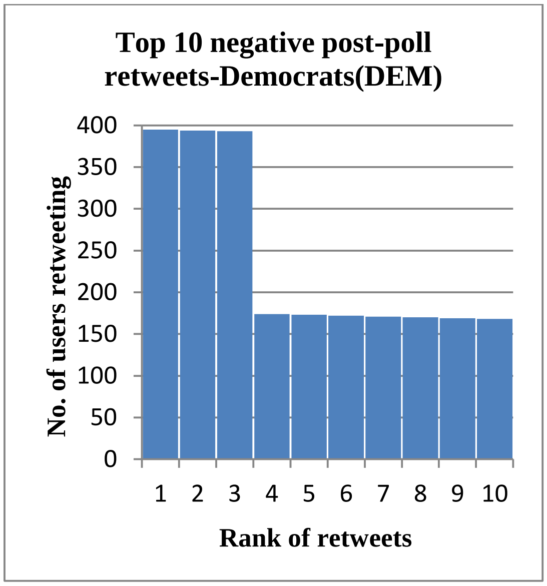


Figure 32(a). Top 10 positive post-poll retweets for DNR.

Figure 32(b). Top 10 negative post-poll retweets for DNR.



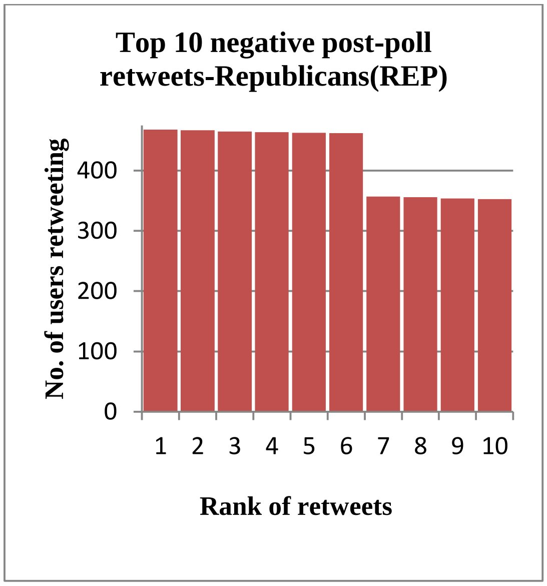


Figure 33 (a). Top 10 negative post-poll retweets for DEM.

Figure 33 (b). Top 10 negative post-poll retweets for REP.

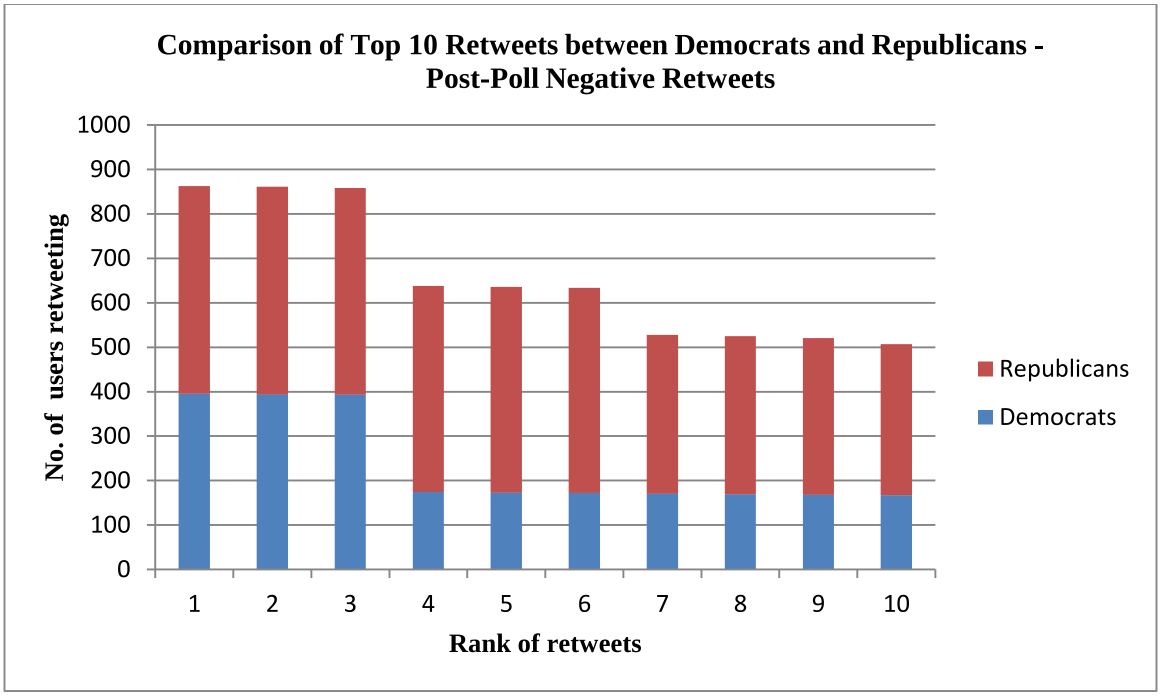


Figure 34. Top 10 negative retweets: DEM and REP – Post-poll.

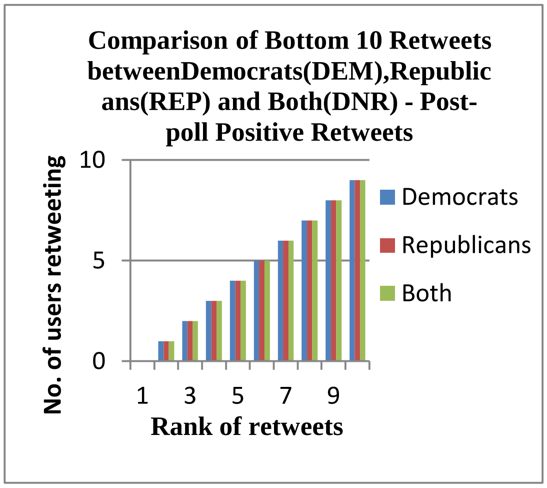


Figure 35(a). Bottom 10 positive retweets:

DEM, REP and DNR – Post-poll.

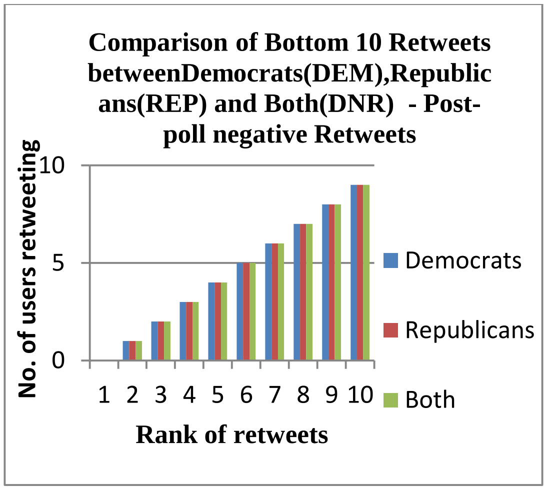


Figure 35(b). Bottom 10 negative retweets:

DEM, REP and DNR – Post-poll.

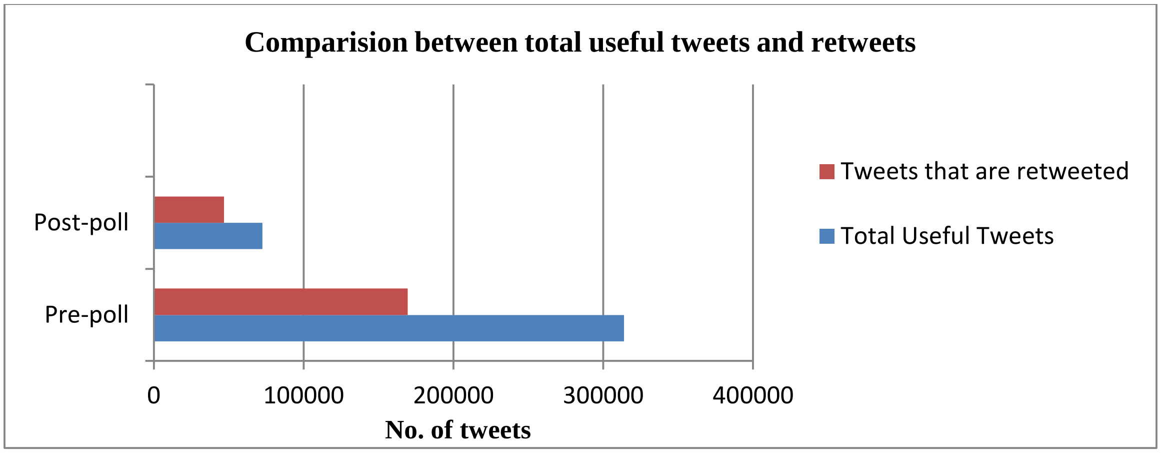


Figure 36. Comparison between total useful tweets and retweets – Pre-poll and Post-poll.

4.3 USERS AND TWEETS

Users and their tweeting behavior are important factors in predicting events based on Twitter data. In the case of election prediction, the volume of tweets that could be associated to a user and a political party signifies the user's interest (positive or negative) in the party. So, we use the number of users and the volume of their tweets as parameters for predictive analysis. We count the users and tweets in all four categories DEM (tweets referring to Democrats), REP (tweets referring to Republicans), DNR (tweets referring to both Democrats and Republicans), and NDR (tweets that do not refer to either party). Tweets associated to distinct users are ranked in descending order. Each tweet count is associated to unique users with that many tweets. Top ten and bottom ten user counts in each category are compared to determine party strength. The comparisons favor the Republicans.

4.3.1 COMPARISON OF TOTAL DISTINCT USERS

Total Number of distinct users for pre-poll (both negative and positive tweets) are 219540 and for post-poll is 47229. Figures 37 and 38 compare the user counts in different groups.

Figure 37 shows distinct users for positive tweets. A comparison between total distinct users for positive tweets in DEM, REP and DNR are shown, REP (Republicans) has more distinct users compared to DEM (Democrats) and DNR (both Republicans and Democrats). REP has 46.8% higher number of users than DEM for pre-poll and 57% more for post-poll. It may be noted that the number of distinct users are more for pre-poll than post-poll for all the three cases indicating that many users stopped tweeting on the topic of election.

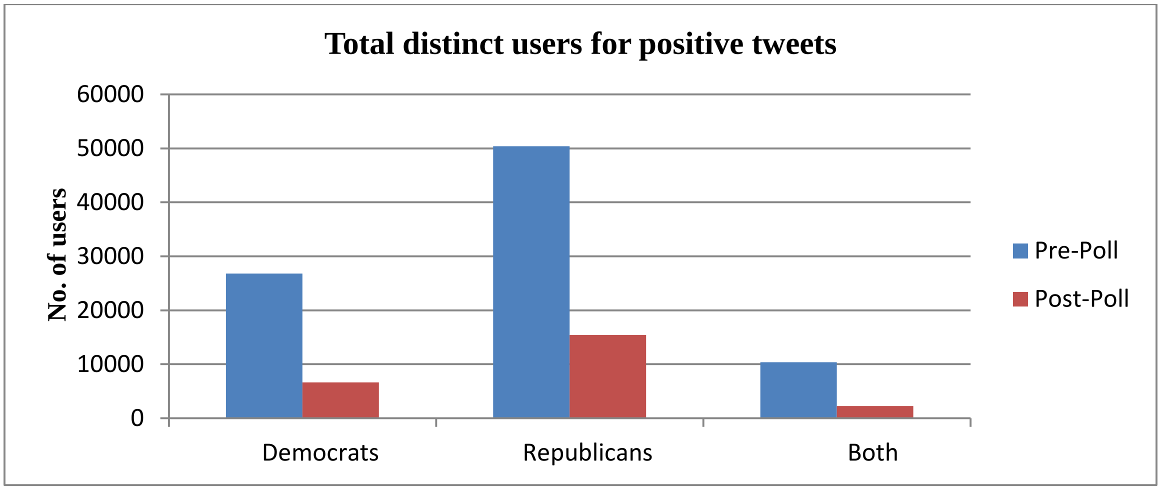


Figure 37. Total distinct users for Positive Tweets.

Figure 38 shows the total number of distinct users for negative tweets whose tweets refer to either Republicans, or Democrats, or Both. The numbers of distinct users are more for pre-poll for all the three cases. The number of distinct users in REP for pre-poll is higher by 20.07% than in DEM. Also the number of distinct users is 38.5% higher for REP than for DEM for post-poll. This clearly indicates that numbers of distinct users are more for REP for both positive and negative tweets.

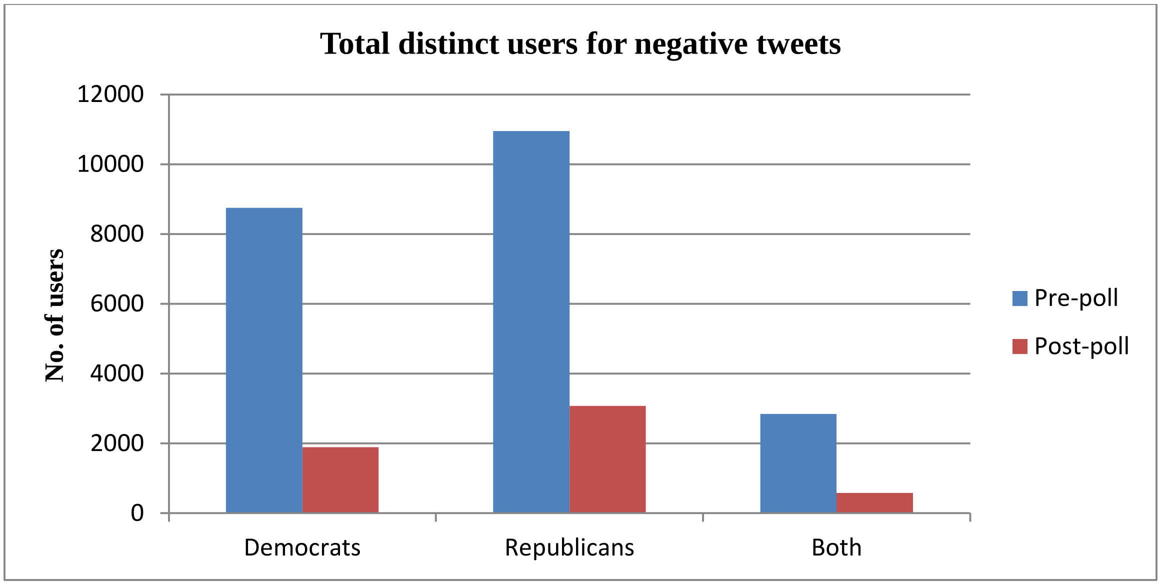


Figure 38. Total distinct users for Negative Tweets.

4.3.2 NUMBER OF TOP 10 AND BOTTOM 10 USERS

In this section we compare the tweet counts in different groups based user rankings. Users are ranked according to their tweet count in each group, DEM, REP, DNR and NDR. A user with highest number of tweets is assigned rank one, next highest rank two and so on. More than one user could have the same rank. Rankings are made for both positive and negative tweets.

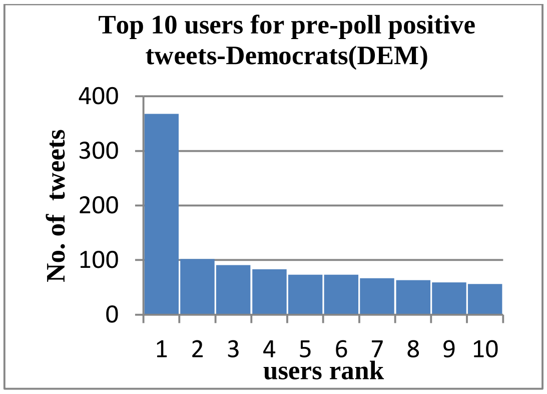


Figure 39(a).Top 10 users for pre-poll positive Tweets - DEM.

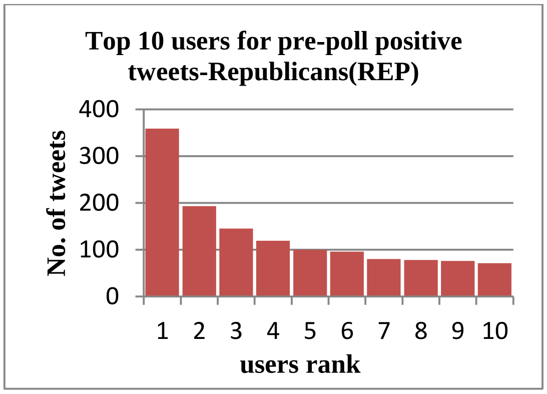


Figure 39(b). Top 10 users for pre-poll positive tweets – REP.

Figure 39(a) shows tweet count of top 10 users for pre-poll positive tweets for DEM. Tweets of a distinct user of first rank is higher by 72% when compared with second rank. Figure 39(b) shows tweet count of top 10 users for pre-poll positive tweets for REP. There is a 46% fall in tweets when first rank user is compared with second rank user. Figure 39(c) shows top 10 users for DNR for pre-poll positive tweets. Percentage decrease of tweets is 70.54% when first and second rank users are compared.

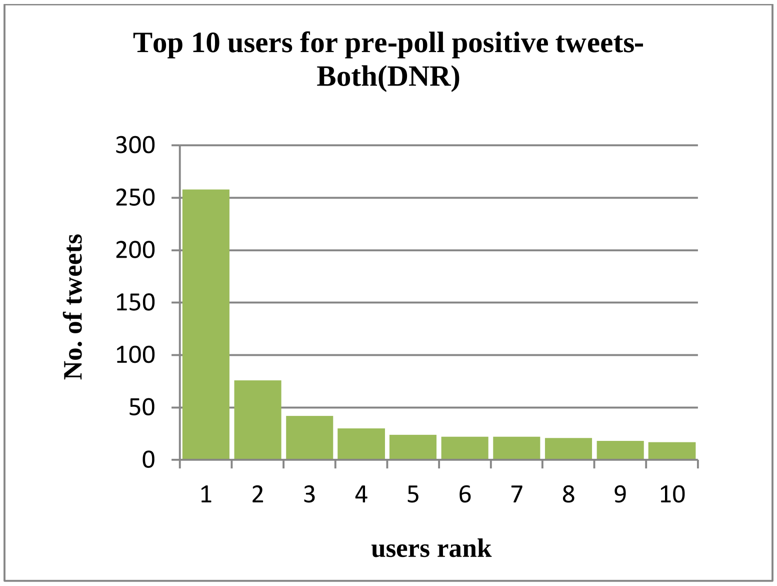


Figure 39(c). Top 10 users for pre-poll positive tweets DNR.

In all three cases, DEM, REP, and DNR, the rank one users tweeted significantly more than the others. One explanation is that these are party affiliated users. In that case, they can bias the tweet count. We propose this problem for future work.

Figure 40 shows the comparison of Top 10 users between the DEM, REP and DNR for pre-poll positive tweet. Except in the case of users in the top rank, REP (Republicans) category has a significant advantage over the DEM (Democrats) category in tweet count. That is, users in the REP category have been tweeting more than the others.

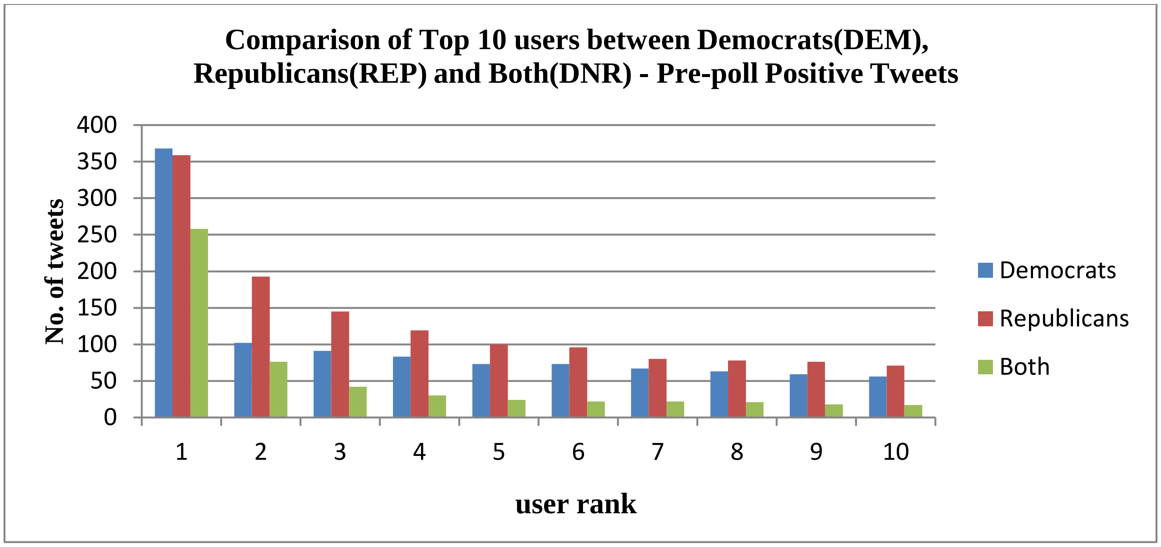


Figure 40. Top 10 users: DEM, REP and DNR - Pre-poll positive tweets.

Figure 41 shows the bottom 10 users for DEM, REP and DNR for pre-poll positive tweets. There is a similar increase in tweets as the user rank increases in all the three cases. This indicates users at the lower rank exhibit the same tweeting behavior as opposed to users at the top rank.

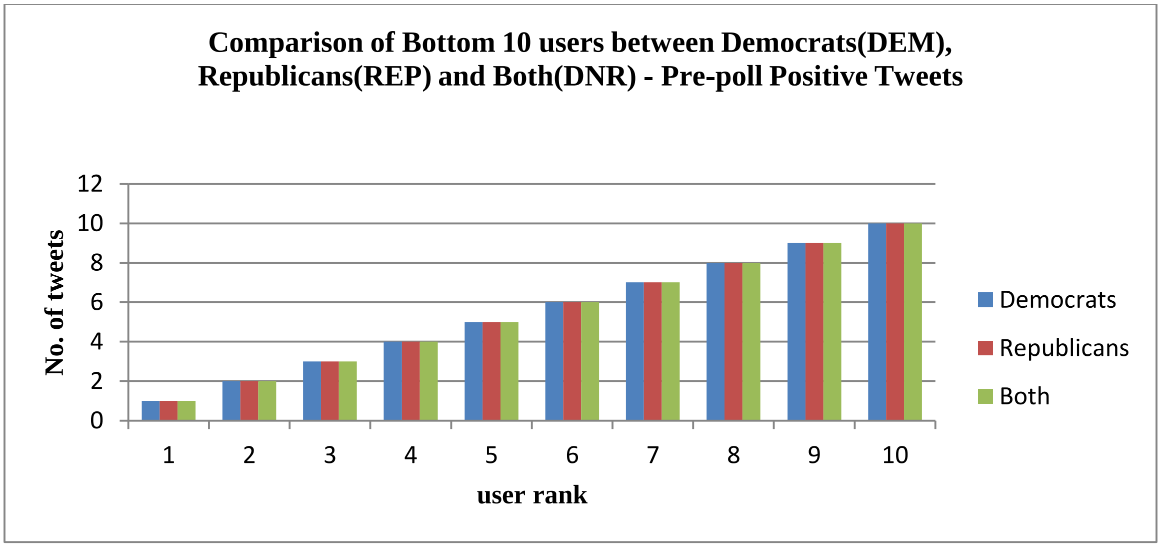
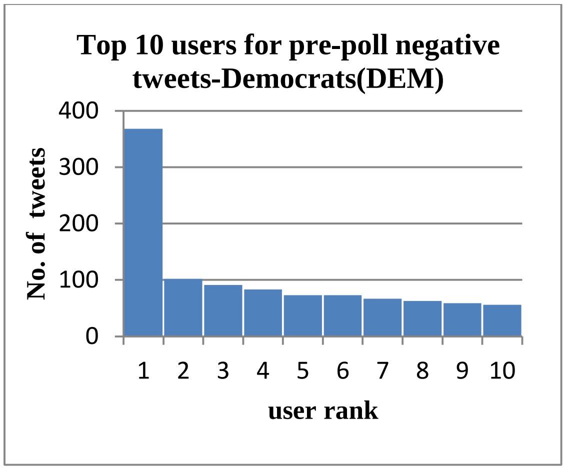


Figure 41. Bottom 10 users: DEM, REP and DNR- Pre-poll positive tweets.



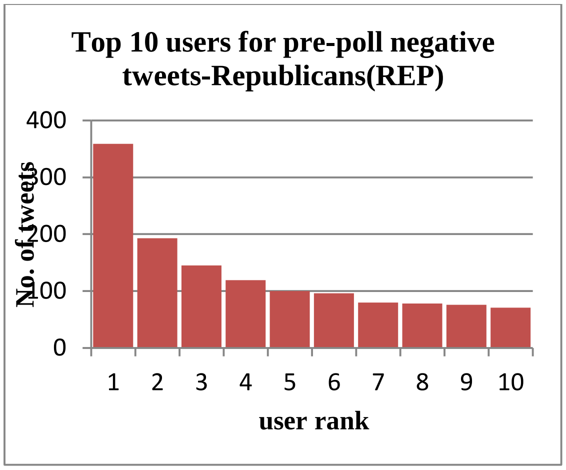


Figure 42(a). Top 10 users for pre-poll negative tweets – DEM.

Figure 42(b). Top 10 users for pre-poll negative tweets – REP.

Figure 42(a) shows Top 10 users for pre-poll for negative tweets in DEM. There is 31.7% fall in the number of tweets from first rank to second rank user and 98.1% fall to tenth rank user. Figure 42(b) shows top 10 users for pre-poll for negative tweet in REP. The fall in tweet count from the first to the second ranked user is not as dramatic as in DEM. Figure 42(c) shows top 10 users for pre-poll for negative tweets in DNR. The number of tweets per user is less when compared to REP and DEM. But, we find that the fall in the number of tweets is 17.4% when the first rank is compared with the second rank.

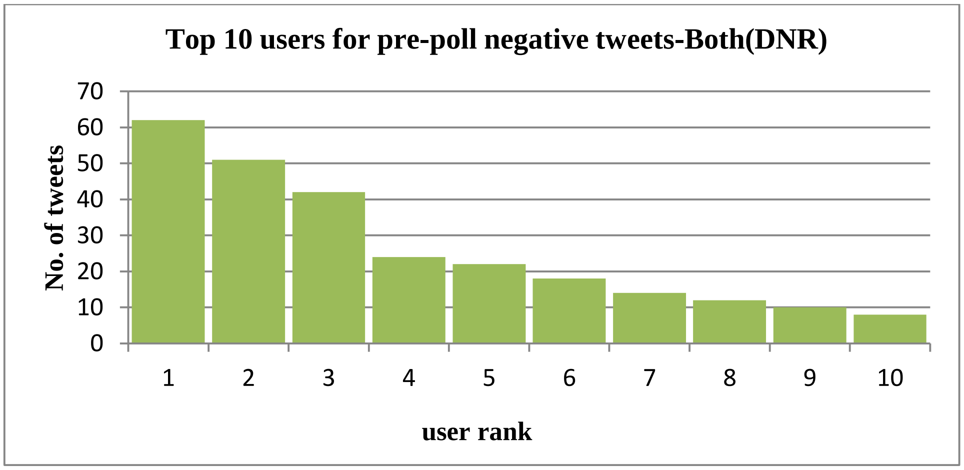


Figure 42(c). Top 10 users for pre-poll negative tweets –DNR.

Figure 43 shows the comparison of top 10 users between DEM, REP and DNR for pre-poll. We find that the number of negative tweets by the top ranked users are more for DEM than REP. This difference very high for the first and second ranked users. In other words, the users who were negative on the parties tweeted extensively, and more in the case of Democrats. As mentioned earlier, this could be party affiliates tweet-bombing the other party. Assuming that is not the case, there are very enthusiastic users supporting the Republican party which can forecast a Republican win.

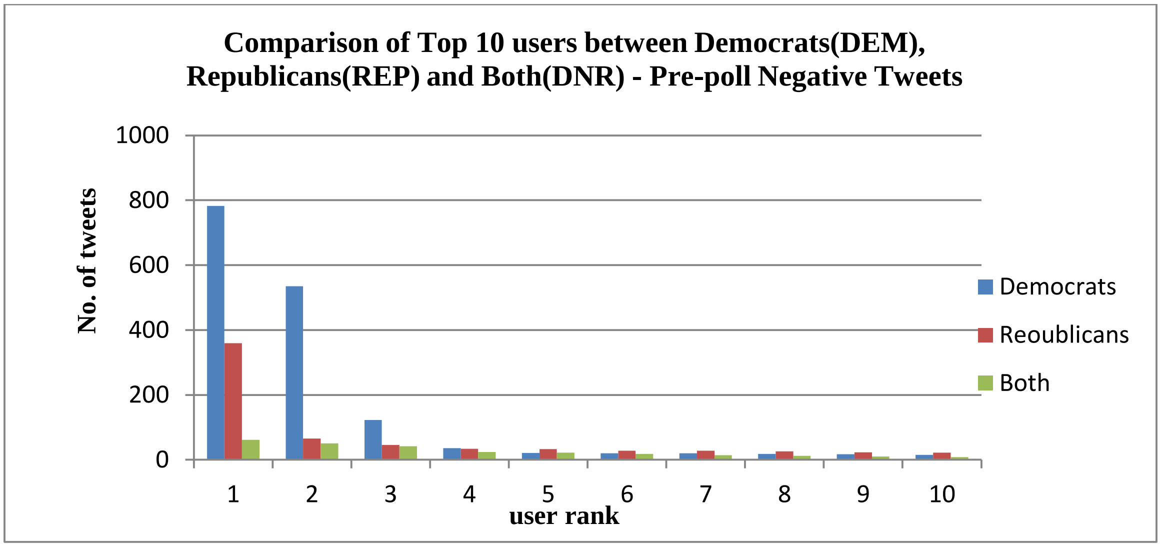


Figure 43. Top 10 users: DEM, REP and DNR) - Pre-poll negative tweets.

The following Figure 44 shows a comparison of bottom 10 users for DEM, REP and DNR. The tweet level at the lower end of the user ranking seems to be the same. There is a linear increase in tweets as user rank increases.

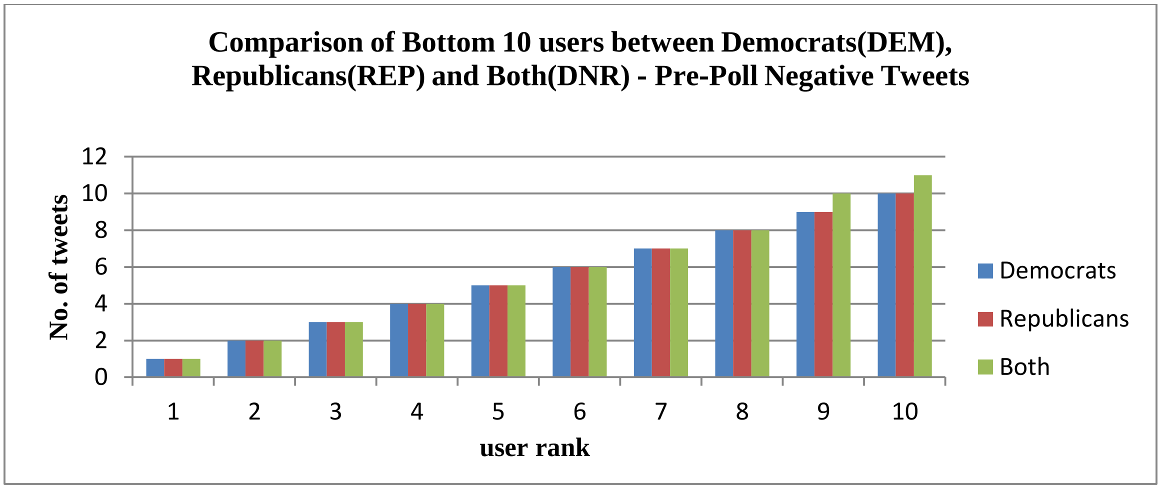
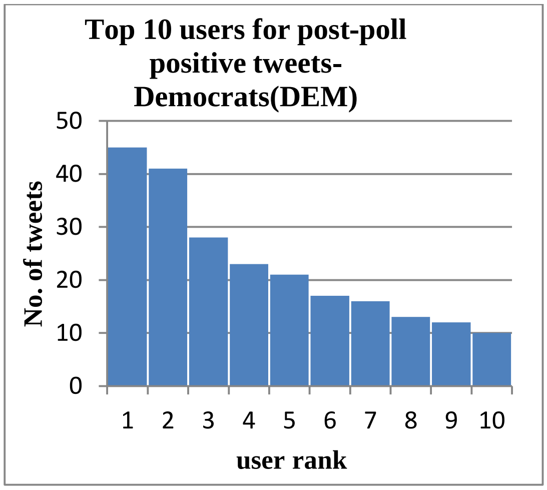


Figure 44. Bottom 10 users between DEM, REP and DNR- Pre-poll negative tweets.

Figure 45(a) and 45(b) show post-poll tweet counts of top ten ranked users in DEM and REP respectively. Tweet counts are much lower than the pre-poll counts which implies that user interest in both political parties changed after the election. Tweet counts decrease much slowly in REP than in DEM. This could be interpreted as the users supporting Republicans are more active than those supporting Democrats.



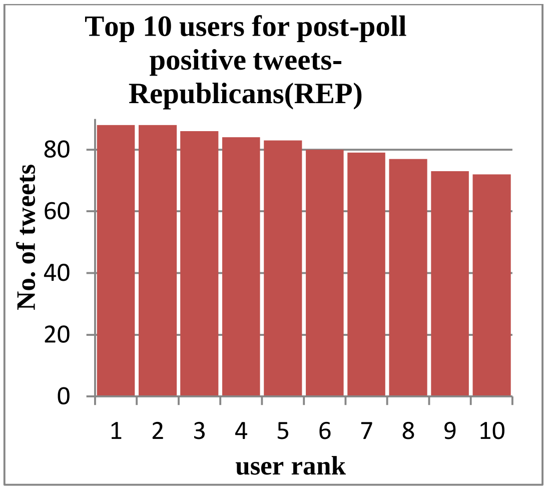


Figure 45(a). Top 10 users for post-poll positive tweets –DEM.

Figure 45(b). Top 10 users for post-poll positive tweets –REP.

Figure 45(c) shows post-poll tweet counts of top ten ranked users in DNR. The numbers of tweets are lesser compare to DEM and REP. There is significant drop in tweet count from second to third ranked user.

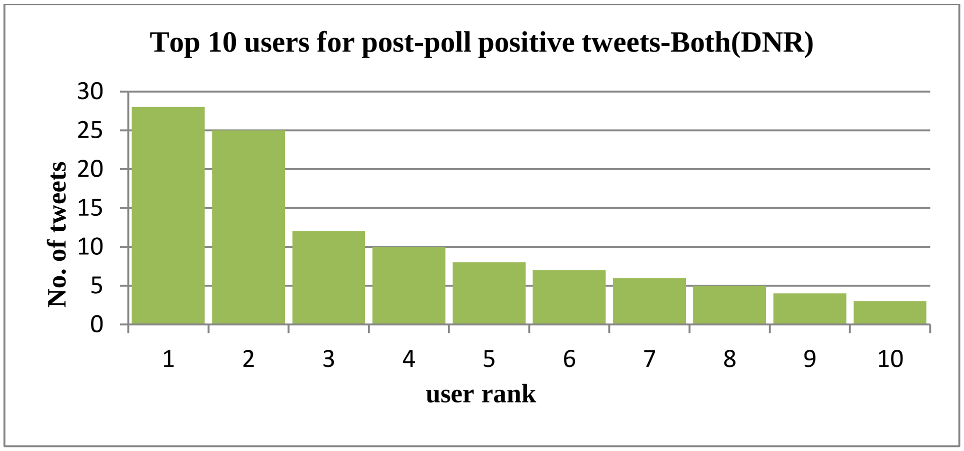


Figure 45(c). Top 10 users for post-poll positive tweets – DNR.

Figure 46 shows the comparison of top 10 users and their tweet count for DEM, REP and DNR for positive tweets. Tweet count for all top ten users in REP are much higher than in DEM.

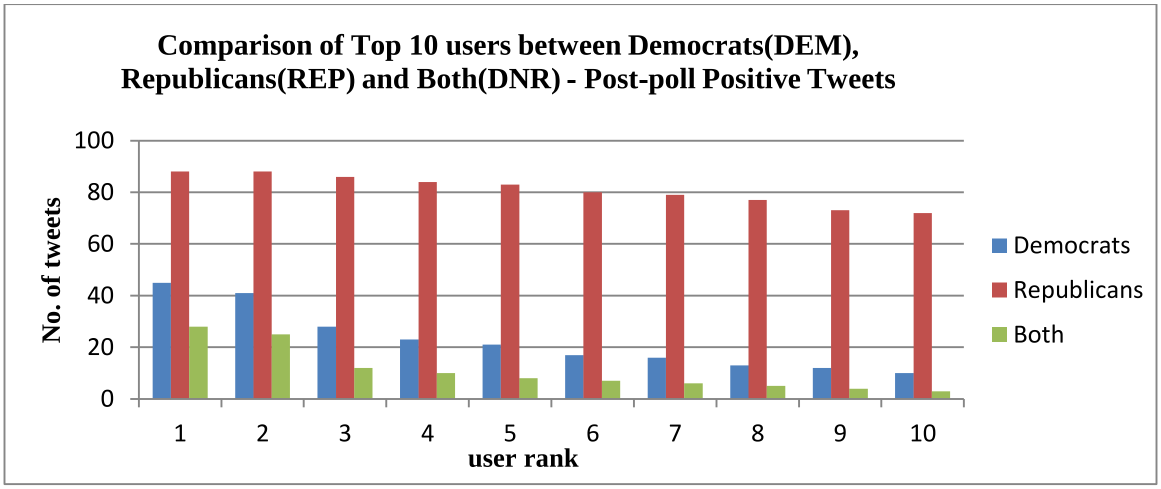


Figure 46. Top 10 users for DEM, REP and DNR -Post-poll positive tweets.

Figure 47 gives the comparison of tweet counts of bottom 10 users for post-poll positive tweets for DEM, REP and DNR. We can see the linear rise in the number of tweets of the users with rank in all three cases.

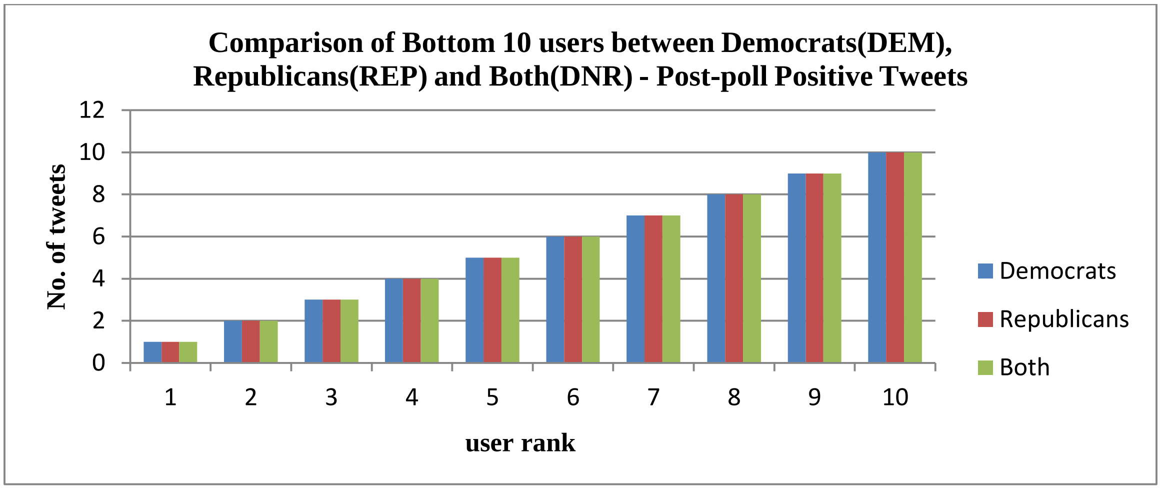


Figure 47. Bottom 10 users: DEM, REP and DNR – Post-poll positive tweets.

Figures 48(a), 48(b) and 48(c) show the tweet counts of the top ten users for post-poll for negative tweets in DEM, REP and DNR. Numbers of tweets are 16, 61 and 81 for the highest ranked user in DEM, REP and DNR respectively.

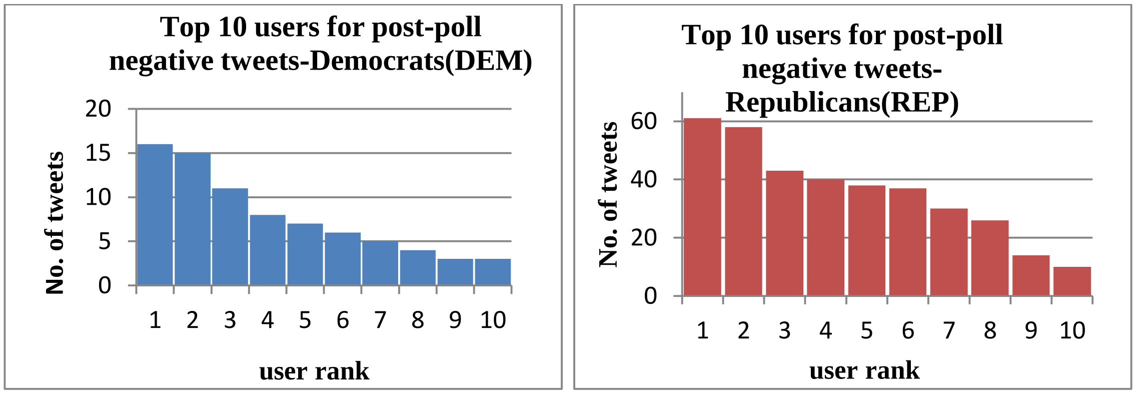


Figure 48(a). Top 10 users for post-poll negative Tweets – DEM.

Figure 48(b). Top 10 users for post-poll tweets – REP.

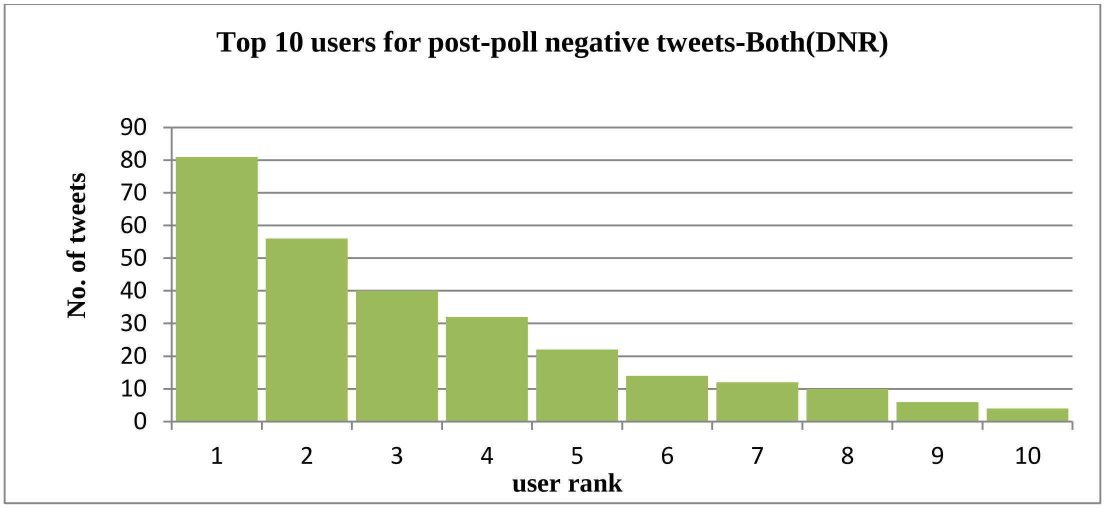


Figure 48(c). Top 10 users for post-poll negative tweets – DNR

Figure 49 shows comparison of top 10 users between DEM, REP and DNR for post poll negative tweets.

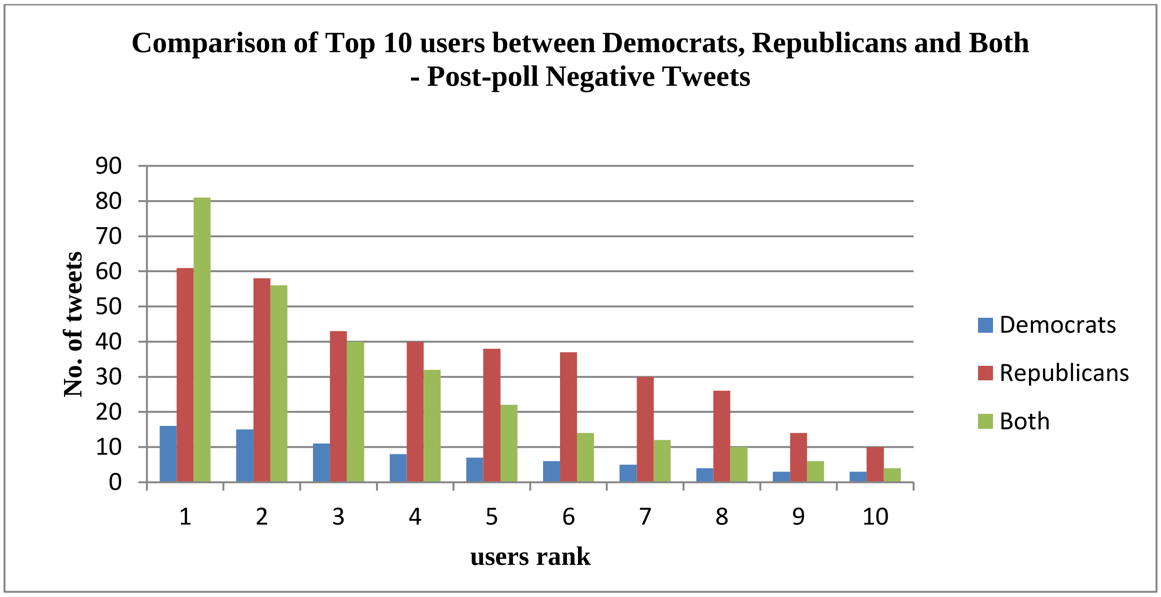


Figure 49. Top 10 users: DEM, REP and DNR - Post poll negative tweets.

Figure 50 shows the bottom 10 users between DEM, REP and DNR for post-poll negative tweets. In all the three cases, there has been gradual increase of tweets of users with rank.

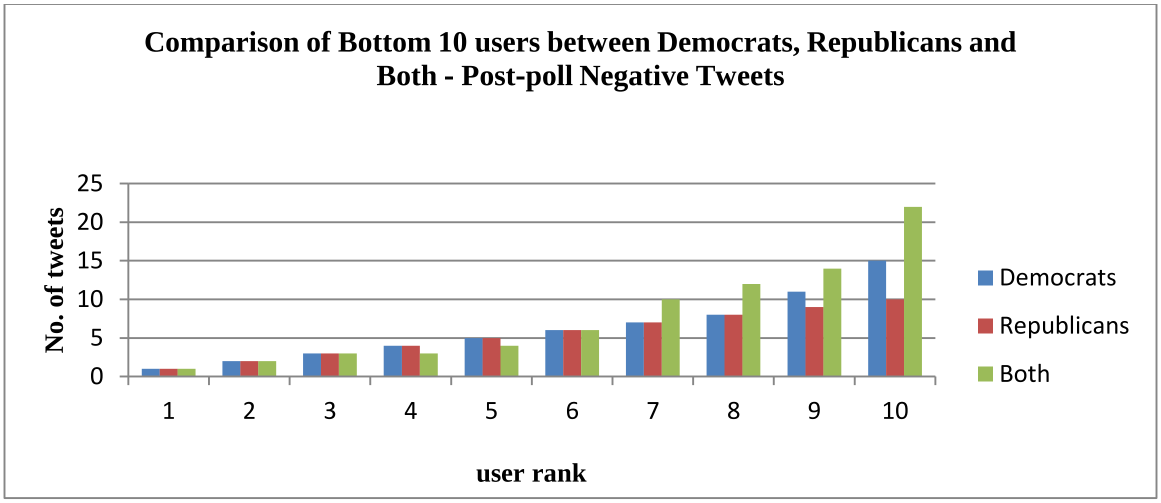
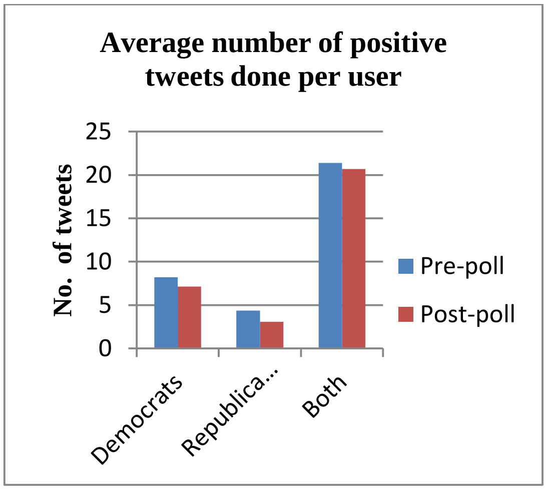


Figure 50. Bottom 10 users: DEM, REP and DNR - Post-poll negative tweets.

4.3.3 AVERAGES NUMBER OF TWEETS DONE PER USERS

Figures 51(a) and Figure 51(b) show the average number of tweets per user for different categories in pre-poll and post-poll data. The average is computed by dividing the number of tweets in each category by the number of users in the corresponding category. There are more tweets per user for Democrats compared to Republicans.



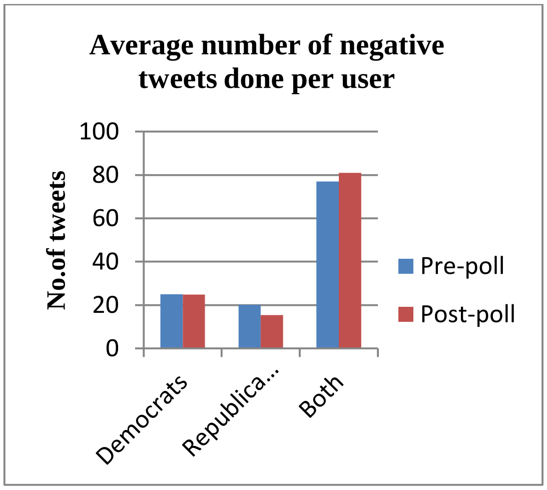


Figure 51(a). Average positive tweets done.

Figure 51(b). Average negative tweets done.

Figure 52 shows the total number of common distinct users for pre-poll and post-poll. It can be seen that the total number of common distinct users are significantly more for pre-poll than for post-poll.

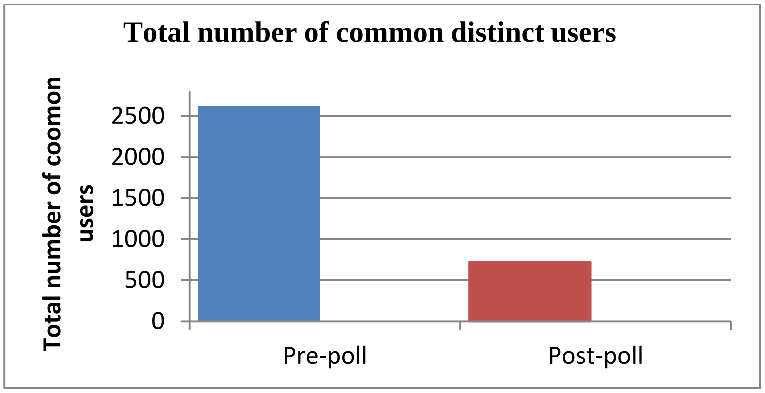


Figure 52. Total number of common distinct users.

4.4 SENTIMENT ANALYSIS

This section is devoted to the discussion of sentiment analysis. Sentiment time series are constructed for all four categories of tweets namely, DEM, REP, DNR and NDR. Each time-series consists of twenty data-points, ten from pre-election and ten from post-election. For each category, positive/negative sentiments is represented by the number of positive/negative tweet for the category. Sentiment is computed based on the keywords and phrases given in the appendices. The graphs in Figures 53(a), 55(a), 57(a), and 59(a) show the positive sentiment graph for the four categories. The graphs in Figures 54(a), 56(a), 58(a), and 59 (a) show the negative sentiment graph for the four categories. Sentiments of all categories are modeled using a Hidden Markov Model (HMM) with two hidden states. Figures 53(b), 54(b), 55(b), 56(b), 57(b), 58(b), 59(b), and 60(b) show the HMM state transition graph of the corresponding sentiment time-series. For our analysis, the HMM states are interpreted as follows:

- State 1 - Not interested
- State 2 – Interested

Positive sentiment associated to Republicans has been consistently in state 2 indicating a win for the Republicans. The states fluctuate after the poll which indicates a change in the interest in the party. On the other hand the state is fluctuating in the case of positive sentiment associated to the Democrats. Negative sentiment has been fluctuating in the case of both parties.

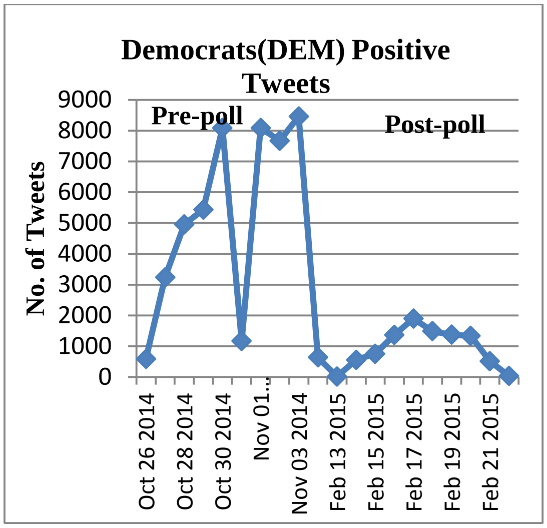


Figure 53 (a). Sentiment analysis for DEM positive tweets.

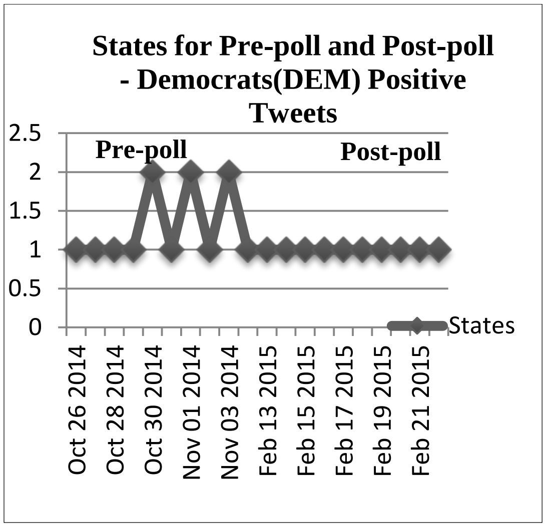


Figure 53 (b). States for pre-poll and post-poll – DEM positive tweets.

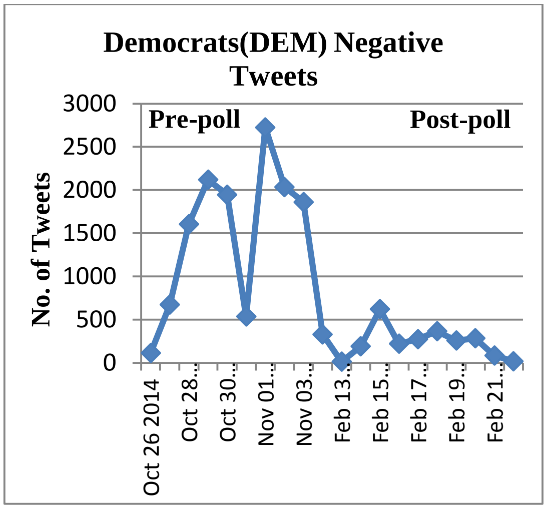


Figure 54(a). Sentiment analysis for DEM negative tweets.

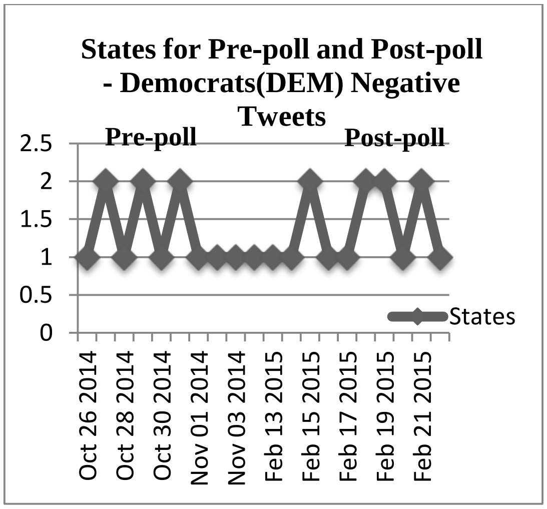
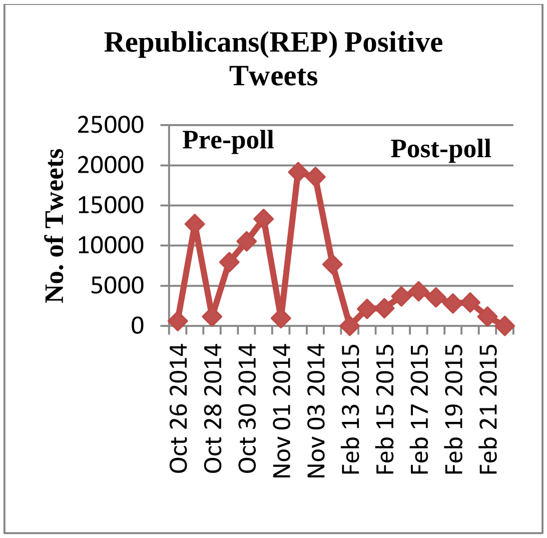


Figure 54(b). States for pre-poll and post-poll – DEM negative tweets.



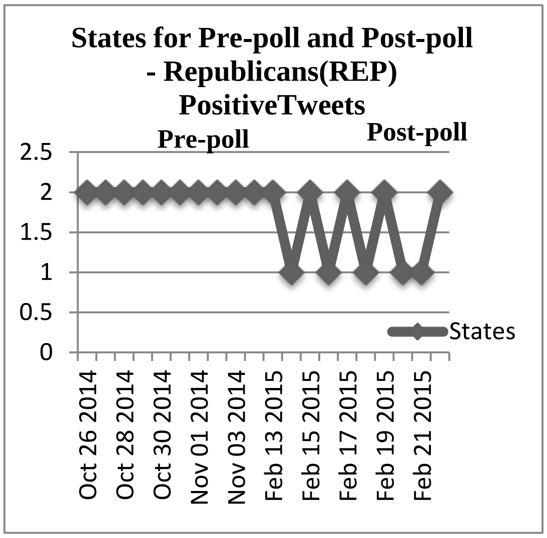
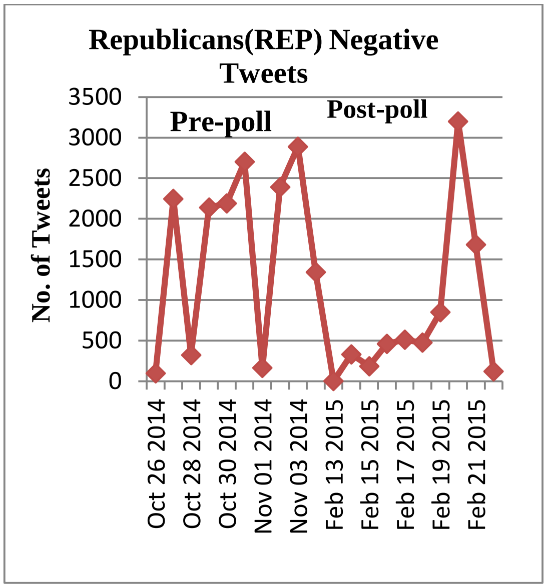


Figure 55 (a). Sentiment analysis for REP positive tweets.

Figure 55 (b). States for pre-poll and post-poll – REP positive tweets.



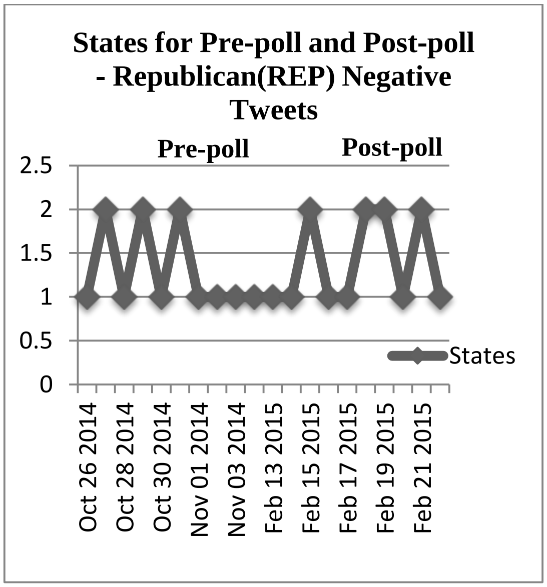
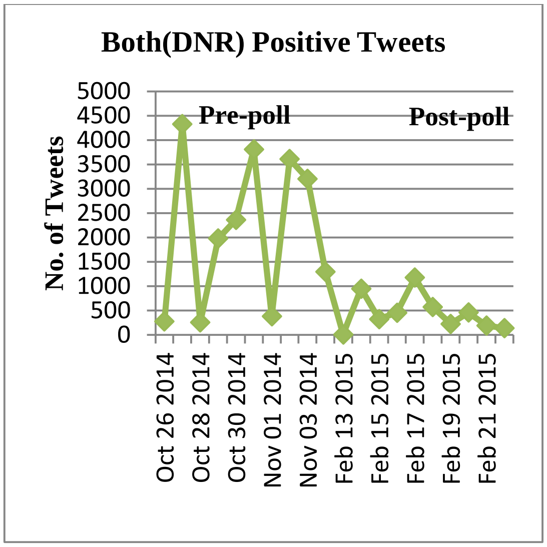


Figure 56 (a). Sentiment analysis for REP negative tweets.

Figure 56 (b). States for pre-poll and post-poll - REP negative tweets.



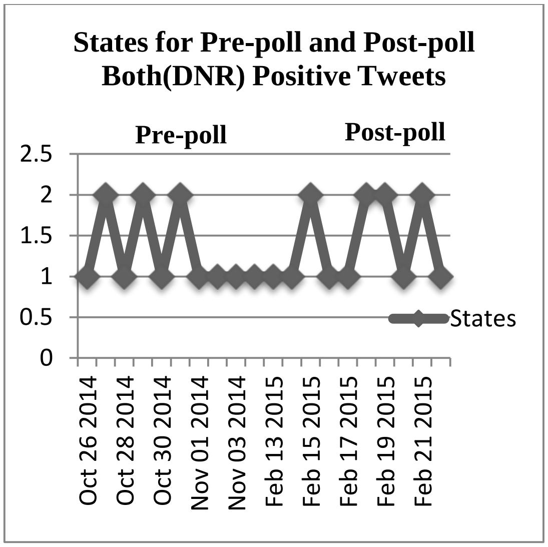
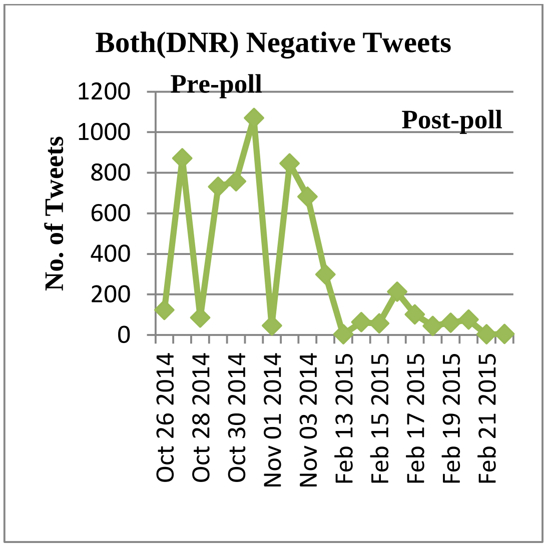


Figure 57 (a). Sentiment analysis for DNR positive tweets.

Figure 57 (b). States for pre-poll and post-poll – DNR positive tweets.



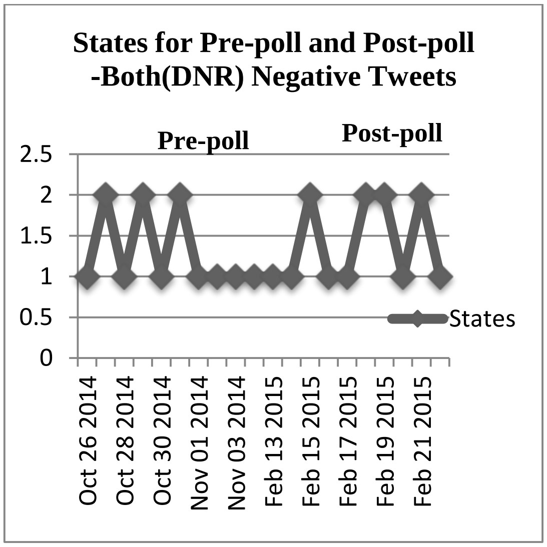
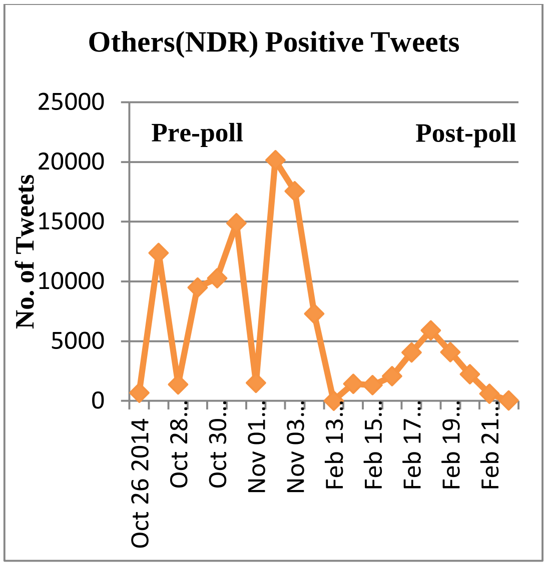


Figure 58 (a). Sentiment analysis for DNR negative tweets.

Figure 58 (b). States for pre-poll and post-poll – DNR negative tweets.



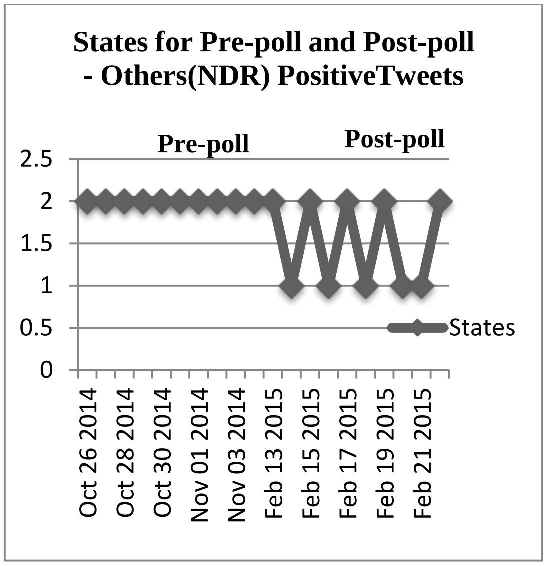
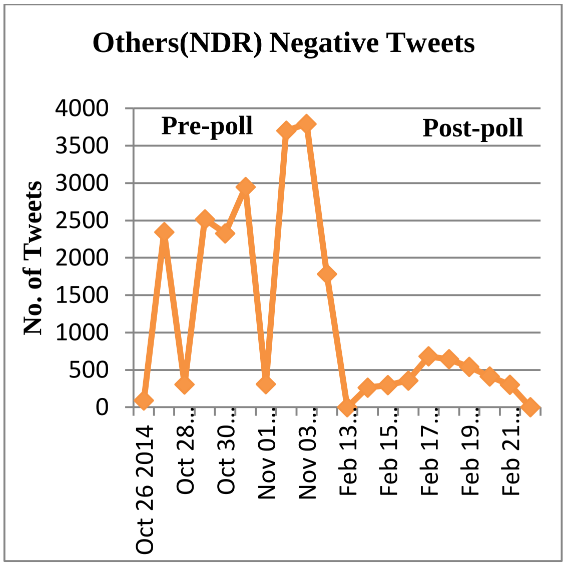


Figure 59 (a). Sentiment analysis for NDR positive tweets.

Figure 59 (b). States for pre-poll and post-poll – NDR positive tweets.



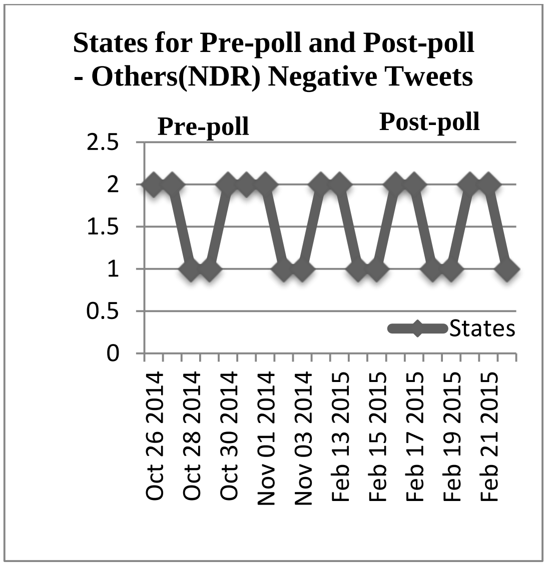


Figure 60 (a).Sentiment analysis for NDR negative tweets.

Figure 60 (b). States for pre-poll and post-poll – NDR negative tweets.

4.5 CORRELATION OF TWEETS AND RETWEETS TO ELECTION RESULT

In this section, we examine whether there is any relationship between the tweets and the number of seats won in the House of Representatives. We compute the ratio R of number of Republican tweets/retweets to the number of Democrat tweets/retweets i.e. R = (number of Republican tweets/retweets) / (number Democrat tweets/retweets). Actual election results show that the ratio of Republican vs Democratic House of Representatives to be 1.31 (= 245/188) which is greater than 1. Table 2 shows the ratio of total tweets and retweets that are significantly higher than 1.31 indicating that tweets and retweets over-estimate the Republican strength in the House.

Table 2. Ratio of Tweets and Retweets.

Total	**Ratio for Republicans/ Democrats**
Tweets	109048/ 62238 = 1.75
Retweets	616843/38123 = 1.62

In order to label the columns of Table 3, let us assume that

1. N1 = Number of positive tweets/retweets for Republicans
2. N2 = Number of negative tweets/retweets for Republicans
3. N3 = Number of positive tweets/retweets for Democrats
4. N4 = Number of negative tweets/retweets for Democrats

(N1-N2)/(N3-N4) is the ratio of the overall sentiment. Table 3 shows the ratios of positive tweets/retweets and negative tweets/retweets. Surprisingly, the ratios of negative tweets are close to the ratio of the actual seats won than the ratio of positive tweets or the ratio of positive sentiments which over-estimate the seats. However these ratios are greater than 1 indicates that the Republicans will have more seats than the Democrats.

Table 3. Predictive Ratio for Positive and Negative –Tweets and Retweets.

Parameters	Positive Tweets	Negative Tweets	Ratio Difference
Tweets	92562/48275= 1.91	16486/13963 = 1.18	76076/34312 =2.21
Retweets	51155/30330= 1.68	10688/7798 = 1.37	40467/22532 =1.79

CHAPTER V

CONCLUSION

The overarching objective of our research is to study the capability of Twitter data as an ex-ante indicator of event outcomes. The 2014 US midterm election has been chosen as the event for this study. This work analyses both Pre-Poll and Post-Poll data from Twitter related to midterm U.S elections. Relevant tweets were extracted from the tweet stream with the help of a Map- Reduce Program in a Hadoop system by specifying an appropriate keywords configuration for running Apache Flume. Several statistics were collected from the extracted tweets.

5.1 PREDICTED AND SENTIMENT CHANGE BASED ON COUNTS

The pre-poll tweet count for the categories of Republicans and Democrat are 35% and 19% respectively. This difference is significant and indicates the Republican Party will be preferred over the Democratic Party. The post-poll tweet count for the same categories are 42% and 16% respectively. It can be clearly seen that people tweeted more about Republicans in both pre-poll and post-poll. This indicates support for the Republicans party did not diminish after the election which they won. In the tweet and user rankings, Republicans were favored at each rank level. The retweet count and the user count support the findings of the tweet count.

5.2 HMM ANALYSIS

HMM is used for modeling shifts in sentiment before and after the event (the poll). Two time-series (positive and negative sentiments) are generated for each category of tweets. These series are used to fit two-state HHMs for all categories. The HMMs show the change in sentiments from pre-poll to post-poll in the cases of DEM and REP categories. There is no noticeable change in sentiment in the case of DNP and NDP categories.

5.3 FUTURE WORK

In this study, we collected data for two 10-day durations. The size of the tweets and duration may not be large enough to generate completely accurate results. Further studies with more extensive data would probably give us better results. These aspects are considered as future research. This study did not have an answer to the question “are the highest ranking users based on tweet count affiliated to the parties?”. This is left as future work.

REFERENCES

1. Hillygus, D. S. (2011). The evolution of election polling in the United States. Public opinion quarterly, 75(5), 962-981.
2. Daniel Gayo-Avello (2011). Don't turn social media into another 'literary digest' poll. Communication of ACM, Vol. 54(10), 121–128.
3. Jihan K. Raoof, Halimah Badioze Zaman, Azlina Ahmad and Ammar Al-Qaraghuli (2013), Using social network systems as a tool for political change, Vol. 8(21), 1143-1148.
4. Shi, Lei and Agarwal, Neeraj and Agrawal, Ankur and Garg, Rahul and Spoelstra, Jacob. (2012), Predicting US primary elections with Twitter URL: http://snap. stanford. edu/social2012/papers/shi. Pdf.
5. Jungherr, A. (2013, October). Tweets and votes, a special relationship: The 2009 federal election in Germany. In Proceedings of the 2nd workshop on Politics, elections and data 5-14. ACM.
6. Castells, M. (2007). Communication, power and counter-power in the network society. International journal of communication, 1(1), 29.
7. Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif (2010): Exploratory Search and Topic Summarization for Twitter. Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC.
8. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010), Election forecasts with Twitter: How 140 characters reflect the political landscape. Social Science Computer Review, 0894439310386557.
9. Daniel Gayo Avello, Panagiotis T. Metaxas, and Eni Mustafaraj. (2010), Limits of electoral predictions using twitter. In Proceedings of 5th ICWSM, AAA Press, 178–185.
10. Johan Bollen, Alberto Pepe, and Huina Mao. (2009), Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. CoRR, abs/0911.1583.
11. O'Connor, B., Stewart, B. M., & Smith, N. A. (2013, August), Learning to Extract International Relations from Political Context. In ACL (1) 1094-1104.
12. Mustafaraj, E., & Metaxas, P. T. (2010), From obscurity to prominence in minutes: Political speech and real-time search.
13. Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. (2011), Understanding the Demographics of Twitter Users. ICWSM 11: 5th.
14. Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wideweb 675-684. ACM.

15. Stirland, S. (2011) Obama's Secret Weapons http://www.wired.com/threatlevel/2008/10/ obamas-secret-w/, Internet, Databases and Psychology, October 29, 2008. Retrieved on August 26, 2011.
16. Drezner, D.W., and Henry F. (2007), Introduction: Blogs, politics and power: a special issue of Public Choice. Public Choice, 134(1-2): 1-13.
17. Livne, Avishay and Simmons, Matthew P and Adar, Eytan and Adamic, Lada (2011), The Party Is Over Here: Structure and Content in the 2010 Election, ICWSM 11, 17-21.
18. Larsson, Anders Olof and Moe, Hallvard. (2012), Studying political microblogging: Twitter users in the 2010 Swedish election campaign, New Media & Society14.5:729-747.
19. Asur S, & Huberman, B. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on Vol. 1,492-499, IEEE.
20. Bermingham, A. and Smeaton, A.F (2011): On using Twitter to monitor political senti- ment and predict election results. In: Proceedings of the Sentiment Analysis where AI meets Psychology Workshop at IJCNLP.
21. Mustafaraj, E. and Finn, S. and Whitlock, C. and Metaxas, P.T. (2011): Vocal minority versus silent majority: Discovering the opionions of the long tail. In: Proceedings of the IEEE 3rd International Confernece on Social Computing, 103–110.
22. Gaurav, M., Srivastava, A., Kumar, A., & Miller, S. (2013, August). Leveraging candidate popularity on Twitter to predict election outcome. In Proceedings of the 7th Workshop on Social Network Mining and Analysis, 7, ACM.
23. Gayo-Avello, Daniel. (2012) , I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper--A Balanced Survey on Election Prediction using Twitter Data." arXiv preprint arXiv:1204.6441.
24. Lewis-Beck, M. S. (2005). Election forecasting: principles and practice. The British Journal of Politics & International Relations, 7(2), 145-164.
25. Gayo Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.
26. Boyd, D, Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on 1-10 IEEE.
27. Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. (2010), What is Twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, 591-600. ACM.
28. Huberman, Bernardo A., Daniel M. Romero, and Fang Wu. (2008), Social networks that matter: Twitter under the microscope.Available at SSRN 1313405, 1-9.
29. Kristina Lerman, Rumi Ghosh, and Tawan Surachawala. (2012) Social contagion: An empirical study of information spread on Digg and Twitter follower graphs, ICWSM, Vol10, 90-97.
30. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.
31. Lampos, V., De Bie, T., & Cristianini, N. (2010). Flu detector-tracking epidemics on Twitter. In Machine Learning and Knowledge Discovery in Databases599-602. Springer Berlin Heidelberg.
32. G. Mishne. (2006), Predicting movie sales from blogger sentimentin In AAAI 2006 sPRING Symposium on Computation Approches of Analysin Wedlogs(AAA1CAAW).
33. Pepe, A., & Bollen, J. (2008, January). Between Conjecture and Memento: Shaping A Collective Emotional Perception of the Future. In AAAI Spring Symposium: Emotion,

Personality, and Social Behavior, 111-116.

34. Livne A., Simmons M. P., Gong W. A., Adar E., Adamic L. A., (2011), Networks and Language in the 2010 Election. 4th Annual Political Networks Conference and Workshops. 1-27.
35. Goldstein, P., and J. Rainey. (2012). The 2010 elections: Twitter isn't a very reliable prediction tool.Retrieved January 10 (2010).
36. Carr. (2010), Facebook, twitter election results prove remarkably accurate,Fast Company, http://bit.ly/dW5gxo.
37. Metaxas, Panagiotis Takis, Eni Mustafaraj, and Daniel Gayo-Avello (2011), How (not) to predict elections. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference, 165-171, IEEE.
38. David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. (2009), Tweet the debates: understanding community annotation of uncollected sources. In Proceedings of the first SIGMM workshop on Social media, WSM’09, 10, New York, NY, USA. ACM.
39. Bollen, J.; Mao, H.; and Zeng, X.-J. 2010, Twitter mood predicts the stock market. Journal of Computational Science, Volume 2, Isuue 1, March 2011, Pages 1-8.
40. Zhang, X., Fuehres, H., & Gloor, P. A. (2011), Predicting stock market indicators through twitter I hope it is not as bad as I fear. Procedia-Social and Behavioral Sciences, 26, 55-62.
41. B. O Connor, R. Balasubramanyan, B. R. Routledge, and N. A.Smith. (2010), From tweets to polls: Linking text sentiment to public opinion time series. In proceedings of 4th ICWSM,AAA Press, pages 122-129.
42. S. Asur, B. A. Huberman, G. Szabo and C. Wang. (2011) Trends in Social Media: Persistence and Decay. In the Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM).
43. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010), Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.
44. Gayo-Avello, Daniel (2012). No, you cannot predict elections with Twitter. Internet Computing, IEEE, 16(6), 91-94. JCR.
45. B. A. Huberman, D. M. Romero, W. Galuba and S. Asur(2011), Influence and Passivity in Social Media.Proceedings of the 20th International Conference on World Wide Web (WWW).
46. Gaffney, Devin. (2010) # iranElection: quantifying online activism.
47. Anders Olof Larsson(2011),Extended infomercialsor politics 2.0 ?: A study of Swedish political party Web sites before, during and after the 2010 election, First Monday Vol. 16, Issure 4.
48. Rumi Ghosh, Kristina Konstantin Voevodski, Lerman Tawan, and Shang hua Teng(2011). Non- conservative diffusion and its application to social network analysis.
49. Golbeck, Jennifer, Justin M. Grimes, and Anthony Rogers.(2010), Twitter use by the US Congress,Journal of the American Society for Information Science and Technology 61, no. 8: 1612-1621.
50. Moe, Hallvard. (2010), Mapping the Norwegian blogosphere: Methodological challenges in internationalizing Internet research.Social Science Computer Review: 0894439310382511.
51. Bruns, Axel, Jean Burgess, Tim Highfield, Lars Kirchhoff, and Thomas Nicolai(2010), Mapping the Australian networked public sphere. Social Science Computer Review: 0894439310382507.
52. Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi.(2010) Measur- ing User Inuence in Twitter: The Million Follower Fallacy. In Fourth International AAAI Conference on Weblogs and Social Media.

53. Guo J, Xu G, Cheng X, Li H.(2009), Named entity recognition in query. SIGIR '09: Proc. of the 32 international ACM SIGIR conference on Research and Development in Information Retrieval. ACM. New York, NY, USA, 267-274.
54. Hua, Wen, Dat T. Huynh, Saeid Hosseini, Jiaheng Lu, and Xiaofang Zhou.(2012), Information extraction from microblogs: A survey. Int. J. Soft. and Informatics6, no. 4 : 495-522.
55. Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., & Aswani, N. (2013, September). TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In RANLP, 83-90.
56. M. van Erp, G. Rizzo, and R. Troncy. (2013), Learning with the Web: Spotting Named Entities on the intersection of NERD and Machine Learning. In Proceedings of the 3rd Workshop on Making Sense of Microposts (#MSM2013).
57. Andreas Jungherr, Pascal Jrgens, and Harald Schoen. (2011), Why the Pirate Party Won the German Election of 2009 or The TroubleWith Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment. Social Science Computer Review. Conference on Weblogs and Social Media (ICWSM) 2011.
58. Ritter, A., Clark, S., & Etzioni, O. (2011, July), Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1524-1534. Association for Computational Linguistics.
59. Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In RANLP, 198-206.
60. Han, B., & Baldwin, T. (2011, June). Lexical normalisation of short text messages: Makn sens a# twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 368-378. Association for Computational Linguistics.
61. Han, B., Cook, P., & Baldwin, T. (2012, July), Automatically constructing a normalisation dictionary for microblogs. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 421-432. Association for Computational Linguistics.
62. Liu, X., Zhou, M., Wei, F., Fu, Z., & Zhou, X. (2012, July) , Joint inference of named entity recognition and normalization for tweets. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 526-535. Association for Computational Linguistics.
63. Peng, M., Huang, J., Fu, H., Zhu, J., Zhou, L., He, Y., & Li, F. (2013) , High Quality Microblog Extraction Based on Multiple Features Fusion and Time-Frequency Transformation. In Web Information Systems Engineering–WISE 2013, 188-201 Springer Berlin Heidelberg.
64. Zhao, Hua, and Qingtian Zeng. (2013), Micro-blog keyword extraction method based on graph model and semantic space. Journal of Multimedia 8, no. 5: 611-617.
65. Polato, I., Ré, R., Goldman, A,, Kon, F. (2014), A comprehensive view of Hadoop research—A systematic literature review Journal of Network and Computer Applications, Volume 46, November 2014,1–25.
66. Dean, J., Ghemawat, S. (2004), MapReduce: simplified data processing on large clusters. Proceedings of the OSDI ’04, Dec. 2004, 137–150.
67. Chia-Wei Lee, Kuang-Yu-Hsiesh, Sun –Yuan Hsiesh, Hung-Chang Hsiao (2014), A Dynamic Data Placement Strategy for Hadoop in Heterogenous Environment, Bid Data Research, Volume 1, August 2014, 14–22.
68. https://dev.twitter.com/streaming/overview
69. https://flume.apache.org/FlumeUserGuide.html

70. http://blog.cloudera.com/blog/2012/10/analyzing-twitter-data-with-hadoop-part-2-gathering-data-with-flume/
71. http://en.wikipedia.org/wiki/JSON
72. http://www.json.org/
73. http://www.copterlabs.com/blog/json-what-it-is-how-it-works-how-to-use-it/

APPENDICES

APPENDIX 1

Keywords for Second Level Filtration

uselection	midterm election	John Jaramillo	Mark Pryor
uselection	midtermelection	Mead Treadwell	Tom Cotton
democrats	Jeff Sessions	Scott Kohlhaas	Mark Swaney
republican	Mark Begich	Sid Hill	Nathan LaFrance
senate	William Bryk	Ted Gianoutsos	Mark Udall
Jaime McMillan	Joe Miller	Zachary Kile	Cory Gardner
Mark Aspiri	Randy Baumgardner	Tom Janich	Bill Hmmons
Raul Acosta	Steve Shogan	Chris Coons	Chris Smink
Kevin Wade	Andrew Groff	Branko Radulovacki	Michelle Nunn
Steen Miles	Todd Robinson	Art Gardner	David Perdue
Derrisck Grayson	Jack Kingston	Karen Handel	Paul Broun
Phil Gingrey	Amanda Swafford	Nels Mitchell	William Bryk
Jeremy Anderson	Jim Risch	Dick Durbin	Doug Truax
Jim Oberwei	Sharon Hansen	Bruce Braley	Ruth Smith
Joni Ernst	Mark Jacobs	Matt Whitaker	Sam Clovis
Scott Schaben	Rick Stewart	Chad Taylor	Patrick Wiesner
Alvin Zahnter	D.J. Smith	Milton Wolf	Pat Roberts
Greg Orman	Randall Batson	Alison Lundergan Grimes	Burrel Charles Farnsley
Greg Leichty	Tom Recktenwaid	Brad Copas	Chris Payne
Matt Bevin	Mitch McConnell	David Patterson	Mike Maggard
Robert Ransdell	Shwana Sterling	Mary Landrieu	Raymond Brown
Vallian Senegal	Wayne Ables	William Waymire Jr.	Bill Classidy
Rob Maness	Thomas Clements	Brannon McMorris	Shenna Bellows
Susan Collins	Ed Markey	Brian Herr	Gary Petrs
Terri Lynn Land	Chris Wahmhoff	Jim Fulner	Paul Marineau
Richard Matkin	Al Franken	Jack Shepard	Sandra Henningsgard
Davd Carlson	Jim Abeler	Mike McFadeen	Ole Savior
Patrick Munro	Heather Johnson	Jack Shepard	Kevin Terell

Stephen Williams	Steve Carlson	Tom Books	Jonathan Rawl
Rex Weathers	William Bond Compton, Jr.	Bill Marcy	Chris McDaniel
Thad Cochran	Thomas Carey	Shawn O's Hara	Amanda Curtis
Dirk Adams	John Walsh	Champ Edmunds	John Bohlinger
Steve Daines	Susan Cundiff	Sam Rankin	Dave Domina
Larry Marvin	Bart McLeay	Ben Sasse	Clifton Johnson
Shane Osborn	Sid Dinsdale	Jim Jenkins	Todd Watson
Jeanne Shaheen	Andy Martin	Bob Heghmann	Bob Smith
Gerard Beloin	Jim Rubens	Mark Farnham	Miroslaw Dziedzic
Robert D'Arcy	Scott Brown	Walter Kelly	Cory Booker
Brian Goldberg	Jeff Beli	Murray Sabrin	Rich Pezzullo
Antonio Sabas	Eugene Lavergne	Hank Schroeder	Jeff Boss
Joe Baratelli	Tom Udall	Allen Weh	David Clements
Ernest Reeves	Kay Hagan	Will Stewart	Alex Bradshaw
Edward Kryn	Greg Brannon	Heather Grant	Jim Synder
Mark Harris	Ted Alexander	Thom Tillis	Sean Haugh
Tim D'Annunzio	Al McAffrey	Constance Johnson	Jim Rogers
Matt Siverstein	Patrick Hyes	Andy Craig	D. Jean McBride-Samuels
Eric McCray	Erick Wyatt	Evelyn Rogers	James Lanford
Jason Weger	Jim Inhofe	Kevin Crow	Eandy Brogdon
Rob Moye	T.W. Shannon	Aaron DeLozier	Joan Farr
Mark Beard	Ray Woods	Jeff Merkley	Pavel Goberman
sWilliam Bryk	Jason Conger	Jo Rae Perkins	Mark Allen Callahan
Monica Wehby	Timothy Crawley	Christina Jean Lugo	James Leuenberger
Jeff Merkley	Mike Montchalin	Jack Reed	Mark Zaccaria
Brad Hutto	Harry Pavilack	Jay Stamper	Joyce Dickerson
Sidney Moore	Benjamin Dunn	Bill Conor	Det Bowers
Eddie McChain	Lee Bright	Lindsey Graham	Nancy Mace
Randall Young	Richard Cash	Tim Scott	Jill Bossi
Thomas Ravenel	Victor Kocher	Rick Weiland	Annette Bodworth
Gordon Howie	Jason Ravnsborg	Larrry Rhoden	Mike Rounds
Stace Nelson	Gordon Howie	Larry Pressler	Gary Davis
Gordon Ball	Larry Crim	Terry Adams	Brenda Lenard
Christian Agnew	Erin Magee	George Filnn Jr.	Joe Carr
John D. King	Lamar Alexander	Ed Gauthier	Joe Wilmoth
Martin Pleasant	Rick Tyler	David Alameel	Harry Kim
Kesha Rogers	Maxey Marie Scher	Michael Fjetland	Wesley Reed
Chris Mapp	Curt Cleaver	Dwayne Stovall	Erick Wyatt
John Cornyn	Ken Cope	Linda Vega	Reid Reasor
Steve Stockman	Emily Marie Sanchez	Jon Roland	Rebecca Paddock
Tanuia Paruchuri	Mark Warner	Anthony DeTora	Charles Moss

Ed Gillespie	Wayshak Hill	Robert Sarvis	David Wamsley
Dennis Melton	Natalie Tennat	Bob Henry Baber	Larry Butcher
Matthew Dodrill	Shelley Morre Capito	Alex Weinstein	Bob Henry Baber
JoHN Buckley	Phil Hudok	Thomas Coyne	Al Hamburg
Chale Hardy	Rex Wilde	William Bryk	Arthur Clifton
Bryan Miller	James Gregory	Mike Enzi	Thomas Bleming
Curt Gottshall	Joe Porambo	Dems	Reps

APPENDIX 2

Negative Keyword and Phrases for Third Level Filtration

Arrogant attitude	Anti social	Anti child	Anti family
Betray	Abuse of power	Assault	Anti job
Abusive	Bizarre	Bureaucracy	Bosses
Consequences	Corrupt	Corruption	criminal
Boastful	Callous	Criminal	Crises
Conspiracy	Controversial	Dissatisfactory	Controversial
De-oriented	Dishonest	Disgustful	Destructive
Disgrace	Excuses	Greed	Failure
Incompetent	Insensitive	Inconsistent	Ill logical
Fictitious	Don’t vote	Pathetic	Pessimistic
Backwardness	Inhumane	Unrealistic	Unreliable
Liars	Stagnation	callous behavior	steal
Scams	Poor performance	Bad reputation	No work
Corruption	Non-approachable	Unpopular	Party corruption
Don’t work unapproachable	Not popular	Scams	Block
Waste	Welfare	Urgent	Problematic
Untruthful	Wrong	Wrongful	Hopeless
Passive	Unauthentic	Unlawful	Unpleasant
Anti - American	Anti – white	Anti –us	Bastards
Loon	Twit	Nitwit	trash
Hippie	Moron	Unthinkable	Fool
Scum	Shill	Hack	Unwisely
Actions	Vague	Zealot	Venomous
Ideologue	Vulgar	Vulnerable	Wasteful
Wasting	Fanatic	Extremist	Nut
Fringe	Wacko	Troll	Loser
Elitist	Doubtful	Exploration	Exorbitant
Extravagant	Failed	Feeble	Hate
Hazardous	Blunder	Helplessness	Bother

Horrified	Baffle	Ill tempered	Horrifying
Imbalance	Impatient	Criticizing	Degenerate
Indecent	Degenerately	Indifferent	Ineffective
Degrade	Dehumanization	Inferior	Inhospitable
Deprived	Depressed	Insupportable	Insincere
Inability	Insane	Diplomatic	Jeopardize
Joblessness	Risk	Threaten	Don’t
Export	Failures	Misuse of funds	Mis-use
Misuse	Arrogant	Misguidance	Miserable
Nasty	Over hipped	Pathetic	Radical
Incapable	Inefficient	Insulting	Vote
Don’t promote	Lagging	Laid off	Irrelevant
Loose	Lost	Lurking	Nonsense
Non confident	Numb	Obstacle	Odd
Offend	Over look	Rival	Adverse
Alarming	Awful	Bad	Callous
Can’t	Clumsy	Detrimental	Dirty
Disgusting	Dishonest	Dishonorable	Don’t
Failure	Ignorant	Immature	Imperfect
Injurious	Pessimistic	Poisonous	Questionable `
Repulsive	Revengeful	Rude	Terrible
Tense	Threatening	Ugly	Terrifying
Unfair	Unfavorable	Unjust	Unpleasant
Unlucky	Unsatisfactory	Unwelcome	Unwise
Contradictory	Decaying	Damaging	Deform
Filthy	Foul	Frightful	Ghastly
Greedy	Horrible	Hostile	Hurtful
Nasty	Negative	Nonsense	Objectionable
Offensive	Oppressive	Ruthless	Scary
Shocking	Sick	Stressful	Stupid
Substandard	Worthless	Abnormal	Aggressive
Aggression	Anarchist	Annoying	Beastly
Blemish	Betrayer	Betraying	Betrayal
Conceited	Not concerned	Conspiracy	Conspirators
Crises	Culprits	Curse	Damaging
Dangerous	Deceiving	Decline	Decrement
Defect	Deficiencies	Deform	Diminish
Degenerate	Degradation	Dejected	Denial
Dislike	Disloyal	Disorientated	Downfall
Exploitation	Extortion	Extravagance	Flaw
Faults	Fool	Foul	Frail
Fraud	Freak	Futile	Frustrated
Non gracious	Humiliating	Harassment	Hardliner
Harmful	Harsh	Immoral	Immature
Heartless	Hateful	Help	Indecent
Indifference	Indifference	Indignity	Ineffective
Inefficient	Unexplainable	Not explainable	Inferior
Intimidating	Loop holes	Jerk	Jobless
Lose	Lunatic	Mad	Low rated performance

Maniac	Mess	Manipulation	Misbehavior
No Action	Misleading	Mockery	Mistrustful
Mistake	Neglected	Obstructing	Obstruct
Offensive	Idiot	Plight	Paranoid
Poverty	Powerless	Refuse	Rough
Rough	Revengeful	Reluctant	Regression
Rigid	Scary	Shameless	Stressful
Substandard	Tarnish	Toil	Timid
Tragic	Bitter	Thrash	Tough
Maker	Troublesome	Unbearable	Unbearable
Unbelievable	Uncertain	Uncivil	Uncivilized
Unconstitutional	Uncontrolled	Unconvincing	Uncooperative
Unemployed	Undignified	Unfair	Unfaithful
Unfriendly	Unhappy	Unhelpful	Unimportant
Unintelligent	Unkind	Unworthy	Washed out
Villains	Villain	Vindictive	Violators
Violator	Waste full	Waste of time	Wretched
Wrong	Wrongful	Wrongfully	Useless
Upsetting	Weird	Worthless	Anti-social
Anti-occupation	Prevent	Cease	Stop
No	No-no	Boycott	Suspicious
Refusal	Horrible	Suppression	Don’t care
Don’t vote for democrats	Don’t vote for republicans	Vote for third party	No love

APPENDIX 3

Positive Keyword and Phrases for Fourth Level Filtration

Accepted	Acclaimed	Achievement	No hate
Good actions	Active	Admire	Agree
Agreed	Amazing	Appealing	Approve
Aptitude	Attractive	Awesome	Beautiful
Beaming	Beneficial	Bliss	Bountiful
Brave	Brilliant	Celebrated	Certain
Champ	Champion	Good choice	Clean
Commendable	Commend	Composed	Congratulation
Constant	Cool	Courageous	Creative
Dazzling	Delightful	Distinguish	Divine
Earnest	Easy	Effective	Efficient
Effortless	Elegant	Enchanting	Energetic
Energized	Enthusiastic	Esteemed	Excellent
Exciting	Exquisite	Fabulous	Fair
Familiar	Famous	Fantastic	No greedy

Fetching	Find	Fitting	Flourishing
Fortunate	Free	Fresh	Friendly
Generous	Genius	Genuine	Giving
Glowing	Gorgeous	Great	Growing
Happy	Harmonious	Healthy	Heartily
Heavenly	Hearty	Heavenly	Honest
Honorable	Honored	Imaginative	Independent
Innovative	Innovative	Instinctive	Intellectual
Intelligent	Invention	Inventive	Jovial
Jubilant	Keen	Kind	Knowing
Knowledgeable	Nice	Novel	Leaned
Lively	Lucky	Lovely	Marvelous
Masterful	Meaningful	Merit	Meritorious
Miraculous	Motivating	Okay	Ok
Open	Optimistic	100 %	One hundred percent
Perfect	Phenomenal	Pleasurable	Pleasant
Poised	Polished	Polished	Popular
Positive	Powerful	Prepared	Principal
Productive	Progressing	Prominent	Protected
Proud	Quality	Quick	Ready
Reassuring	Refined	Refreshing	Rejoice
Reliable	Remarkable	Respected	Restored
Reward	Rewarding	Rewarded	Proud
Positive	Safe	Satisfactory	Secure
Simple	Smile	Skillful	Soulful
Sparkling	Special	Spirited	Stunning
Successful	Sunny	Superb	Simple
Supporting	Surprising	Terrific	Thorough
Thrilling	Trusting	Truthful	Unwavering
Upright	Upstanding	Valued	Vibrant
Vitreous	Victory	Vigorous	Virtuous
Vital	Welcome	Well	Whole
Whole sum	Willing	Wonderful	Worthy
Wow	Zeal	Active	Caring
Actively	Care	Caring	Confident
Control	Courage	Cares	Challenge
Challenges	Confident	Control	Duty
Empowerment	Hard-work	Humane	Mobilized
Good moral	Opportunity	Passionate	Peace
Pioneer	Principled	Principle	Unique
Success	Truth	Admirable	Admire
Adore	Advanced	Agree	All around
Ambitious	Amazing	Appealing	Applaud
Approved	Appreciate	Attentive	Awarded
Balance	Believable	Best known	Best performing
Best selling	Better known	Blissful	Bloom
Breath-taking	Brave	Brainy	Better than
Better than	Breakthrough	Bright	Brightest

expected			
Clear-cut	Compatible	Brilliant	Calm
Capable	Celebrated	Cherish	Chivalrous
Civilized	Clarity	Clean	Operative
Correct	Cost-effective	Cost saving	Courageous
Creative	Daring	Credible	Decisive
Decent	Dignified	Deserving	Distinguished
Durable	Divine	Eager	Earnest
Economical	Educated	Effectiveness	Elegance
Eminence	Enhance	Enhancement	Enlighten
Enterprising	Enlightenment	Entertaining	Enthusiastic
Exceptionally well	Exceeding well	Extraordinary	Exquisite
Fabulous	Fair	Faithful	Famous
Fearless	Favorite	Favorable	Flawless
Flourishing	Fluent	Fortunate	Fortunately
Friendly	Fulfillment	Generous	Gifted
Glory	Good	Goodness	Goodwill
Gorgeous	Graceful	Graceful	Grand
Hardworking	High spirited	Honest	Hopeful
Humble	Ideal	Events	Impartial
Inexpensive	Invaluable	Keen	Justify
Joy	Lawful	Liberate	Lovable
Low cost	Low risk	Loyal	Magnificent
Meritorious	Mind blowing	Mind – blowing	Modest
Modem	Neat	Nice	Non violence
Non – violence	Noteworthy	Open	Optimist
Outperform	Outstanding performance	Outstanding	Outshine
Over joyed	Overtook	Paramount	Passionate
Patient	Peace	Passionate	Patient
Peace	Patriotic	Perfect	Perfection
Preferable	Prestigious	Problem free	Productive
Progress	Promising	Promoter	Reasonable
Qualified	Prosper	Prosperous	Proper
Record setting	Refine	Reforming	Remarkable
Respectful	Responsibly	Salute	Satisfactory
Satisfied	Self-determination	Self determination	Result satisfaction
Sensation	Self sufficient	Self sufficient	Self respect
Sensible	Sharp	Significant	Simplest
Skilled	Speedy	Stable	Straight
Forward	Streamline	Strong	Stunning
Talented	Successful	Superb	Supportive
Thoughtful	Vote for republican	Vote for democrats	Undisputed
Unquestionable	Unselfish	Versatile	Valuable
Victorious	Well balanced	Well behaved	Well educated
Well established	Well informed	Well regarded	Well managed
Well mannered	Well wisher	Willing	Willing
Support whole	Winnable	Winner	Wise

heartedly			
Wonder	Wonderful	Workable	World famous
Worthwhile	“ Don’t let that happen”	“Make tit happen”	The Republican party has a message
Win Fights	I love this	I would vote	Takes to be
Are pledging to win	#VoteMatters	You’re Born Support	Talks ups
Today is the anniversary	Yes I would turn	Enough to vote	Including 14 Republicans
I may Run for	Don’t let it happen	Make it happen	I’m Republican Running
Republicans think Mitt Romney should	Taking over the younger votes	Eye big gains	Go to vote
Republicans take big lead	Please go vote the Republican tickets	Republican president is the best	Vote is crucial
Vixctories in U.S in midterm elections	Wave in the midterm election	Voters trust	Approach democracy
Enough to vote	You’re helping	Yes I would turn	Talks
Must stage a show	Going to retake	Gets elected	Letter to senate
You should see	I wanted to just vote	Enough to vote	You’re helping

APPENDIX 4

Steps for Installing Hadoop and Flume

To install Hadoop on OS X , we followed the following steps were:-

1. Creating a designated Hadoop user on the system.
2. Install/Configure preliminary Software
3. Setting up Remote Desktop and Enabling Self Login
4. Downloading and Installing Hadoop
5. Formatting and Running Hadoop
6. Stopping the Hadoop DFS

To install Hadoop on MAC System, it is necessary to have JAVA and SSH installed the system. It is preferred that one must upgrade the JAVA for the latest version. SSH is Pre-installed in the system however it may be enabled too. Next step is to download

Latest version of Hadoop from internet. Hadoop is unpacked in the directory of choice and ownership permission for the directory in set.

Second step is to configure Hadoop- There are two files which are to be modified when configure Hadoop. The first is conf/hadoop-env.sh. The next part is to set Hadoop-site xml. Set Hadoop.tmp.dir(which should be sent to the directory of your choice) and mapred.tasktracker.maximum properly to the file. This will effectively set the maximum number of task that can simultaneously run by the task tracker. The last step involves formatting the namenode and the testing system and this will give the output. To run hadoop we start DFS, it will start up, a Task Tracker, Job Tracker and DataNode on the machine. To stop the Hadoop, run the stop all.sh.command.

Steps to execute the Hadoop program:

1. Go to hadoop directory
2. bin/hadoop namenode -format
3. sbin/start-dfs.sh
4. sbin/stop-dfs.sh

Apache Hadoop has overcome its initial unstable phase and now has grown into solid and stable stage. Though Hadoop System were designed to optimize the performance of large batch jobs, but recently the number of applications are increasing. There is a rising demand for Sharing Hadoop clusters which leads to increasing system heterogeneity [1,4].

We conclude that Apache Hadoop is efficient, robust, reliable and scalable framework to store, process, transform and extract big data in cluster of nodes.

APPENDIX 5

List of Initial Probabilities assumed for each category:

Type of Tweet Count (Pre-Poll and Post-Poll)	Probabilities of Transition Vector	Probabilities of Transition Vector
Democrats – Positive	[0.5, 0.5 0.5,0.5]	[0.1, 0.1, 0.1, 0.2, 0.5 0.1, 0.1, 0.1, 0.1, 0.6]
Democrats – Negative	[0.5, 0.5 0.5,0.5]	[0.3, 0.4, 0.3, 0.1, 0.5, 0.5]
Republicans – Positive	[0.5, 0.5 0.7,0.3]	[0.1, 0.4, 0.1, 0.4; 0.1, 0.5, 0.1, 0.3]
Republicans – Negative	[0.8, 0.2 0.7,0.3]	[0.3, 0.2, 0.3, 0.2; 0.1,0.1,0.1,0.7]
Both – Positive	[0.7, 0.3 0.7,0.3]	[0.1, 0.1, 0.1, 0.2, 0.5 0.1, 0.1, 0.2, 0.1, 0.5]
Both – Negative	[0.7, 0.3 0.7,0.3]	[0.1, 0.1, 0.1, 0.2, 0.5 0.1, 0.1, 0.1, 0.1, 0.6]

Others– Positive	[0.5, 0.5 0.5,0.5]	[0.1, 0.1, 0.6 0.2; 0.1, 0.5, 0.1, 0.3]
Others– Negative	[0.5, 0.5 0.5,0.5]	[0.1, 0.1, 0.6 0.2; 0.1, 0.5, 0.1, 0.3]

VITA

Zenia Arora

Candidate for the Degree of

Master of Science

Thesis: PREDICTION OF U.S ELECTION USING TWITTER DATA: A CASE STUDY

Major Field: COMPUTER SCIENCE

Biographical:

Education:

Completed the requirements for the Master of Science in your Computer Science at Oklahoma State University, Stillwater, Oklahoma in July, 2015.

Completed the requirements for the Bachelor of Technology in Computer Science and Engineering at M.D University, Delhi, India in 2015.

Experience:

- Successfully completed Internship/Industrial Training Project for 8th Semester from Persistent System Limited, Pune. India from 28th Jan 2013 to 30th May 2013 which is global company in partnership with Cisco, IBM, Microsoft, Nokia, and specializes in Software Product and technology innovations.

- Successfully completed training from IBM Pvt. Ltd. as a summer Intern in 2012. The name of the project was "**Password Management**". The project has been done using PL/SQL on maintaining the security system of the database so that no unauthorized user could access the account and misuse the data. The project was based on changing the password after every month so that the data in the record could be kept safe. A notification was sent to the user regarding the change of password (i.e. 5-7 days before the expiry date).

- Successfully completed training from IBM Pvt. Ltd. as a summer Intern in 2011. The name of the project was "**Rational Tools**". The aim of the project was to have the basic understanding of Rational Software and Rational Tools. Rational software helps to drive greater values from your software investments and deliver innovative products and services. The project dealt with some of the rational tools used by Bharti Airtel (Telecom firm in India) such as clear case, clear quest, rational robot, rational performance tester, and rational functionality tester.