Prediction and Analysis of Pakistan Election 2013 based on Sentiment Analysis

By

Muhammad Asif Razzaq

2010-NUST-MS PhD-IT-13

Supervisor

Dr. Ali Mustafa Qamar

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(May 2014)

Approval

It is certified that the contents and form of the thesis entitled "**Prediction and Analysis of Pakistan Election 2013 based on Sentiment Analysis**" submitted by **Muhammad Asif Razzaq** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Ali Mustafa Qamar**

Signature: _____
Date: _____

Committee Member 1: **Dr. Muhammad Murtaza Khan**

Signature: _____
Date: _____

Committee Member 2: **Dr. Muhammad Muddassir Malik**

Signature: _____
Date: _____

Committee Member 3: **Hafiz Syed Muhammad Bilal Ali**

Signature: _____
Date: _____

I dedicate this work to my beloved father (late), mother for her unconditional love and family for their help & support in my life.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at National University of Sciences & Technology (NUST) School of Electrical Engineering & Computer Science (SEECS) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Muhammad Asif Razzaq**

Signature: _____

Acknowledgment

Up and above everything all Glory to **ALMIGHTY ALLAH**, The Beneficent, The most Merciful and Most Compassionate. It's a great blessing from Almighty Allah that gave me the health and strength to do this research work and fulfill my dreams. Without His Blessing nothing could have been possible.

I have been fortunate and privileged to discuss and share views with different fellow research scholars in Machine Learning Group at SEECS. They have provided me with invaluable knowledge and have shown me what it means to perform careful and responsible research.

I would also like to pay heartfelt thanks to supervisor, **Dr. Ali Mustafa Qamar** for his expert guidance and Committee members including Dr. Muhammad Murtaza Khan, Dr. Muhammad Muddassir Malik and Hafiz Syed Muhammad Bilal Ali for their unconditional support and inspirations for me during different research phases. Despite all assistance provided by supervisors and others, I alone remain responsible for any error or omissions which may unwittingly remain.

Muhammad Asif Razzaq

Abstract

The significance of social media has already been proven in provoking transformation of public opinion for developed countries in improving democratic process of elections. On the contrary, developing countries lacking basic necessities of life possess monopolistic electoral system. In this system, candidates are elected based on tribes, family backgrounds, or landlord influences. They extort voters to cast votes against their promises for the provision of basic needs. Similarly voters also poll votes for personal interests being unaware of party manifesto or national interest. These issues can be addressed by social media, resulting as ongoing process of improvement for presently adopted electoral procedures. On the brighter side, people of Pakistan utilized social media to garner support and campaign for political parties in General Elections 2013. Political leaders, parties, and people of Pakistan empowered themselves additionally with social media, in disseminating party's agenda and advocacy of party's ideology on Twitter without much campaigning cost. To study effectiveness of social media inferred from individual's political behavior, large scale analysis, sentiment detection & tweet classification was done in order to classify, predict and forecast election results. The experimental results depicts that social media content can be used as an effective indicator for capturing political behaviors of different parties Positive, negative and neutral behavior of the party followers as well as party's campaign impact can be predicted from the analysis. The analytical findings proved to be having considerable correspondence with actual results as published by Election Commission of Pakistan.

Contents

List of Abbreviations

Abbreviations	Descriptions
ECP	Election Commission of Pakistan
PTI	Pakistan Tehreek-i-Insaf
PMLN	Pakistan Muslim League Nawaz
PPP(P)	Pakistan People's Party Parliamentarian
PAT	Pakistan Awami Tehreek
IND	Independent
Pos	Positive
Neg	Negative
Neu	Neutral
MQM	Mohajar Qaumi Movement
KPK	Khyber PakhtunKhawa
NB	Naïve Bayes
KNN	k Nearest Neighbor
RF	Random Forest
NBMN	Naïve Multi-Nominal
Prind	Probabilistic Indexing
SVM	Support Vector Machines
BOW	Bag of Words
JSON	JavaScript Object Notation
MAE	Mean Absolute Error
MAsen10	2010 US Special Senate elections in Massachusetts
USsen10	US Congressional Elections 2010

List of Figures

List of Tables

Chapter 1

Introduction

In virtual environment, interaction of people from different geographical regions through which they communicate is the fundamental motif behind social media. In Section 1.1, importance of social media and microblogging is outlined. In Section 1.2, societal impact of twitter is stated. Section 1.3 deals with motivation behind this development of thesis. Section 1.4 identifies thesis objectives, how this study was carried out. Section 1.5 summarizes the twitter's political contribution made during this study. Finally Section 1.6 outlines how different chapters are organized in the rest of thesis.

1.1 Importance of Social Media and Microblogging

The dramatic increase of social media application was seen in recent years. Social media comprises of many internet-based applications crafted for the creation and exchange of user based contents [28]. Individuals, organizations and different communities are benefited from social media applications. By the passage of time, computer based social media applications were remodeled as 'micro-blogs' for hand-held devices like mobiles. These micro-blogs are responsible for exchanging text, image and media links among users. Popularity for micro-blogging is rising by each coming day. People need to keep themselves in touch with others to be better informed. This explosive growth is very much seen generally for Facebook, LinkedIn, FriendFeed, MySpace etc., and especially Twitter.

1.2 Societal impact

Our current focus is 'Twitter', a popular micro-blogging site, with rich contents of user created data comprising of text, media and images URLs. Twitter has millions of users, generating contents periodically, called 'status messages'. These contents as in Fig. 1.1 are available publically, in most of cases utilized by researchers for the purpose of opinion-mining, sentiment analysis, predictions and finding masses attitudes and trends . As already said, User based textual contents are readily available and can be obtained by using APIs. These APIs cater for researchers in obtaining real-time as well as some third party APIs also give access to crawl historical twitter data. Once data is collected, it would undergo several validated steps before classification, required for opinion-mining and sentiment prediction. The results obtained will pave us in predicting future for certain brands, movies, polls, public moods, and also for many other individuals, organizations and community's objectives. Popularity of movies and forecasting box-office revenues was discussed in depth by Asur and Huberman [23]. In political context, predicting German Elections 2009 [3], United States Senate Elections 2010 [15][25], Singapore General Elections 2011 [2] using twitter was studied and sentiment analysis was proven as valid indicator in deducing results in contrary to original polls [3]. Public moods were studied using twitter feeds during Arab Spring' in which social media especially Twitter got special attention and usage. It started from self-immolation of Tunisian which triggered Arab World, resulting in ouster of Tunisian President. These fires lead dethronement of Egyptian President Hosni Mubarik and later on collapse of Gaddafi's Libyan regime. Upon studying the protest wave, one cannot deny the influence of social media igniting transformational power and bottom-up mobilization of citizens. This learning and awareness changed the lives of Arabs for their advancement democratically. Despite all pros and cons social media applications have the overall impact over societal improvement. A virtual crowd-sourcing society is formed over the internet sharing impartial viewpoints access to each other which increases cooperation and collaboration. We can say it is not a prerequisite for nation development but it can contribute significantly in purification of nation building processes. Social media play an important role in shaping political debates in US, even debates are designed on the outcomes of social media demands [5].

Figure 1.1: Word Cloud for Social Media

1.3 Motivation

My own personal motivations for embarking on this Research Area were to further educate myself on Tweet Classification, to learn more about twitter framework, and to provide and utilize different classification techniques. I have long been interested in Machine Learning methods for various purposes, and I enjoy working with tweets and their relation with natural languages. Doing corpus-based work in Natural Language Processing and reading the literature on the topic always makes me excited to work.

When I started work on my own Tweet Classification, I found various freely and customizable available tools for my use. These tools were from Python & Perl. Starting from Tweet collection till classification whole exercise helped me in polishing my abilities and developing a strong contemporary application focusing on predicting and analysis of Pakistan's General Elections 2013 results.

1.4 Thesis Objectives

Starting with the assumption, percentage of tweets volume received is directly proportional to size of party, no of candidates and number of constituencies which the party covers for contesting elections. Larger the party, it would attract larger number of twitter users similarly smaller parties will have little presence of followers on Twitter. Using tweets with clear-cut little content information, we would like to investigate whether the above assumption is enough for party popularity and can it be used to predict election results comparable with actual General Elections results at the National level and Provisional level. Tweets mentioning party names, party leader names are

reflective of vote bank in actual or not. We would like to investigate political sentiment analysis with bottom-up approach considering individual political affiliation to the party level campaign sentiment.

1.5 Thesis Contribution

Collection and study of Tweets related to Pakistan's different political parties was one of the unique and wonderful experiences for me. This corpus was later on undergone through different phases, like cleaning, annotation, feature selection and finally classification resulting in sentiment prediction. The research performed in this area proved to be solely. First time ever in the history of Pakistan social media, a model was built to discuss and evaluate people's participation over Twitter on National and Provincial level. My focus of study is replicating the previous knowledge in this field and to extend it with respect to Pakistan's General Elections 2013. Following are the compiled research questions to be answered.

- How Pakistan Election campaign was expressed on Twitter?
- How accurately we can predict election results from National level to Province level based on Twitter.
- Detect causes if any, for sentiment shift during Pre-Election Campaign and their effects using Twitter.
- Effect of social media on twitter users for developing countries like Pakistan.
- Influence of Twitter users with respect to election from outside the Pakistan.

1.6 Thesis Organization

The rest of thesis is organized in the following Chapters, with Chapter 2 explaining background of twitter, how it works? what is its role in politics? Is there any contribution of Twitter in Pakistan's Politics. Also a brief of history of Pakistan's elections is discussed. In Chapter 3, details are given for past work done as literature review discussing various important papers discovering role of Twitter in campaigns for elections. Later on discussions

are made for Twitter, whether it can be used as political deliberation or not based on facts collected from Literature Review. Also this chapter explores well known classification techniques, pre-processing methods, feature selection techniques and evaluation measures. Chapter 4 describes methodologies adopted for sentiment prediction, selection of model and algorithms. A detailed analysis for both qualitative and quantitative is carried out in this chapter. Chapter 5 discusses experimental results and outcomes. Deliberation about neutral tweets in addition to positive and negative tweets have been made in chapter, as this information goes waste mostly if done binary sentiment classification.

Chapter 2

Twitter Overview & Political Contributions

This chapter comprises of various sections with each relating importance of twitter and its impact in politics. In Section 2.1, popularity of Twitter among social media applications is outlined. Section 2.2 discusses societal contribution of twitter in the field of politics is stated. An introduction is given how Twitter can be used generally for election campaign. In Section 2.3, use of Twitter in Pakistan's General Election 2013 is highlighted. Section 2.4 summarizes background of previous general elections of Pakistan. In Section 2.5 discussion is made, whether opinions identified from tweets can be used to predict elections results.

2.1 How Twitter works?

Twitter a social networking service and micro-blogging service was launched in 2006 and gained popularity among people with over 500 million registered users (source:*www.wikipedia.com*). Twitter is in the top ten most used social media applications on the internet. Twitter simply broadcast tweets with 140 characters or less to the followers (one who chooses others tweets to be posted in their time-line) around the network. These tweets represent any information in text form or shared link relating personal activity, entertainment, sports, or politics.

Nowadays importance of Twitter has risen as it has proved to be fast, low cost medium for disseminating real time information for any sporting event, disaster, academic conference, political activities, elections etc. Twitter was

found to be most effective during some of the worth mentioning events like Michael Jackson's Death 2009, Iranian election protest 2009-10, FIFA World Cup 2010, NBA Finals 2010, Egyptian revolution 2011, Sandy Hurricane 2012 and US Election 2010 [13]. During these events twitter users updated their statuses and tweeted thousands of tweets per second and setting up tweet propagation per unit time records. These tweets resulted in revolutions for certain nations [1] [30], sometimes celebrating one's win, sometimes benefiting people suffering from disasters like earthquake, flooding, and sometimes aiding politician for their campaigns. These events got popularity on the basis of terms, words or phrases referred by users commonly in their tweets, so most frequent used words would rather help in trend topic setting. These trending topics would quickly suggest for a user "what is happening right now". These trending topics mostly range between real-time events. People send their reactions, opinions, endorse someone else ideas and discuss about these events in the form of tweets [22]. These status updates are reflected on user's page and also in the time-line of follower's pages. Direct messages can be sent to targeted users using '@username' for one-to-one correspondence. Tweets starting by 'RT' is indicator for re-tweet in which user endorse and propagate someone else's interesting post to rest of his followers. These tweet statuses and re-tweets will form trends if more frequently discussed over twitter.

2.2 Twitter & Politics

Microblogging applications enables user to share their views publically. It is being believed and proved that microblogging services such as twitter have the potential for increasing political participation [5]. Campaigns for elections have found totally new platform provided by different social media applications. The popularity rise can be studied for developed and underdeveloped countries. Political parties, party workers, and politician post their campaign messages over fan pages and twitter accounts, maintained by themselves or paid party employees. Number of followers reflects the popularity of political figure and agenda as a simple rule. Conventional campaigns for elections involves cost, time, exertion for dispersing ones motivational opinion in order to get attention of voters, in contrary to this, social media campaign especially twitter involves no such cost. The politicians and political parties can outreach voters free of cost and in no time in deliberating party's agenda and ideology. The political parties can address twitter users from all walks of life, mostly involving younger generation by disseminating appealing

agenda goals and objectives. While campaigning on twitter, political parties projects their positivity, in the meantime also propagates negativity of opponents. This sets a mobilization mechanism for twitter users and followers to vote.

2.3 Twitter and Pakistan

We have seen large amount of contributions investigating relations between twitter and politics for developed countries. They have matured their democratic electoral processes and procedures. The causes behind elections are interests of National level for developed nations. In case of Pakistan with low literacy rate and majority population living in rural areas without basic civic facilities like gas, electricity, health facilities. These lacking facilities have strong influence over voter's mind during elections. Since 1947 Pakistan's majority elections ended up by electing people with feudalism background. It is very difficult for a common people to participate and win elections. Provision of civic facilities acts as a major manifesto for politicians. Voters too cast votes by electing faces for their personal interests like people who promised with provision of drinking water, gas and other civic needs.

With the effect of globalization, freedom of press, awareness through social media networking, People of Pakistan openly participated in election campaign using several social media applications. Taking twitter in consideration, political parties created their party pages, most of political leaders used twitter for their social campaign. Large number of people from different cities of Pakistan followed them. They shared their views in shape of tweets regarding political leadership, parties' & their agenda, demands & desires pre-elections and post-elections. These online discussions in the form of tweets motivated younger generation in particular to actively participate in election campaign and ending up by casting their votes for the desired candidates.

We have seen a positive attitude of people connecting each other from different parts of Pakistan in order to empower change with collaboration by actuating younger literate generation into democratic electoral system. Some of the top Hashtags associated with Pakistan General Elections 2013 are shown in Fig. 2.1

Figure 2.1: Hashtags associated with Pakistan General Elections 2013

2.4 Background for Previous Pakistan's Election & Details

Having free, fair and transparent elections is an integral part of electoral system of any democratic country. Pakistan held several General Elections since 1947, involving mobilization of whole nation, political parties, leadership and people. People elected members for National Assembly and Provisional Assemblies by casting their votes. Since 1970 Pakistan had around 10 General elections [27]. Fortunately General Elections 2013 were held under the rule of law for the first time as a result, first time in history of Pakistan a smooth transition of Government took place in a democratic manner.

2.5 Twitter as Prediction Tool

In 2009 American Association of Public Opinion Researchers (AAPOR) spent $2 billion for online research, out of which $1.7 billion were spent for traditional forums, weblogs, and political discussion boards. The main focus of AAPOR was contents obtained from online surveys are beneficial and reliable rather than user generated contents. On the other hand Tamasjan worked around the point that Micro-blogging tweet contents obtained from twitter shows close correspondence with parties' and politician's posi-

tions. They demonstrated that offline political landscape can be visualized through contents of twitter messages. [3]. Their investigation proved that Micro-blogging content's information do base upon opinions of certain users or vetting of opinions of reliable users in their social networks. Thus their opinion on Micro-blogging forum may have certain weight which cannot be ignored and can be negotiated for opinion mining and predictions. Tamsjan further proved that accurate results and prediction for elections outcomes based on the data obtained from social media rather than political forums, and weblogs already supported by AAPOR. Social media also possess same features pertaining to data as of obtained from financial market containing aggregation mechanism from dispersed bits of information. Price system information can be referred to Micro-blogging twitter data by considering the size of followers and re-tweet rates, and most frequently used terms. These features influenced by human behavior can be used to predict outcomes for currently occurring events [15]. In addition to above mentioned salient features, detected sentiment also holds valuable data which be aggregated in to meaningful, predictable information. Previous studies as of 2010 US Elections [13], Singapore elections [1] with extensive statistical outcomes proved twitter to be a social sensor for the prediction of electoral results which can be related to poll results.

Chapter 3

Literature Review

Extensive research has been carried out on Twitter data during the recent years. Tweets lie in text category and in order to classify text, this chapter explores prominent work done so far in predicting election results based on twitter dataset. Section 3.1, discusses various contributions made by different researchers for the prediction of election results for different countries and comparisons with actual results. Section 3.2 & 3.3 unveils state of the art definition for text classification. In Section 3.4 generic steps to perform text classification are discussed. Section 3.5 reveals about different classifiers used for experiments in this thesis and finally in Section 3.7 evaluation technique of precision and recall is discussed.

3.1 Significant Contributions

The following subsections discusses Twitter contributions for German Elections 2009, Singapore General Elections 2011, 2010 US Special Senate elections in Massachusetts, US Presidential Elections 2008, Irish General Elections 2011, US Midterm Elections 2010, and finally comparison of French and US Presidential Elections 2012.

3.1.1 How Twitter affected German Federal Elections 2009?

The political developments and events are reflected in tweets and even resulted in top trends maintained by twitter. German Federal Elections were held in 2009, Tumasjan revealed tweets obtained for related political parties, leader asserted resemblances with originally compiled results and concluded

Party	All mentions		Election	
	Number of tweets	Share of Twitter traffic	Election result*	Prediction error
CDU	30,886	30.1%	29.0%	1.0%
CSU	5,748	5.6%	6.9%	1.3%
SPD	27,356	26.6%	24.5%	2.2%
FDP	17,737	17.3%	15.5%	1.7%
LINKE	12,689	12.4%	12.7%	0.3%
Grüne	8,250	8.0%	11.4%	3.3%
			MAE:	1.65%

** Adjusted to reflect only the 6 main parties*

Figure 3.1: Share of Tweets and election results [3]

twitter can act as a mirror to offline political landscape [3]. They downloaded tweets in German language and translated them for their onwards analysis. They investigated ideological ties between parties and political coalitions by studying tweets. Besides this they also compared share of tweets with actual votes for 6 main parties with calculation of MAE as 1.65% as shown in Fig. 3.1[3]. They also evaluated tweets containing multiple mentions later on bonding in political ties for different parties. Tweet messages suggested that they are not only used for spreading political opinions but also discuss those opinions. Tamasjan also found a relationship amongst sentiment profiles of politician and parties in context with elections campaign. While considering participation of Twitter users they concluded that in their collected tweet sample, heavy users were unable to politicize the discussion through their sentiment. On the other hand Tamasjan admits that tweets sample may have not been representative of German electorate. According to them, quite number of tweets might have remained unaddressed as they only collected tweets containing politician and political parties' names, ignoring all other tweets & replies. According to Tumasjan they considered all tweets as one document for sentiment detection, which was a drawback in their research by ignoring once again individual tweet's sentiment, which would have different impact for tweet classification. Their research work hovered around the mentions in the tweets, without considering any plausible sentiment finding for tweets.

Party	% tweets	% votes	% error
WP	20.83 (2)	12.83 (2)	8.00
SPP	4.41 (6)	3.11 (6)	1.30
SDP	11.07 (4)	4.83 (4)	6.24
SDA	1.81 (7)	2.78 (7)	-0.97
RP	5.22 (5)	4.28 (5)	0.94
NSP	13.86 (3)	12.04 (3)	1.82
PAP*	42.8 (1)	60.14 (1)	-17.34
MAE			5.23

Note. Numbers in brackets indicate relative rank
*Ruling party. MAE = mean absolute error.

Figure 3.2: Percentage of Tweets and Actual Votes [1]

3.1.2 Twitter and Singapore General Elections 2011

The importance of twitter and other social media applications was also unveiled during Singapore General Elections 2011 which found twitter to be integral part of election campaign and mobilization of citizens to cast vote [1]. Singapore follows Westminster electoral system with two electoral divisions, Single Member Constituencies (single member is elected) and Group Member Constituencies (Group of members elected). Marko worked around Twitter messages mentioning political parties and their candidates' names. They tried to prove relation of tweets mentioning party names and their candidates with their respective share of the vote at the National level. They focused on seven political parties and calculated MAE of prediction as 5.23% whereas on constituency level they were unable to find convincing correlation between percentage of the votes for opposition and tweets received by them as given in Fig. 3.2[1]. They only focused tweets containing mentions, ignoring tweet text, which could have played more significant role in finding out sentiments for a political party and candidates. This detected sentiment would have shed more light in reading out the minds of voter about the political parties and candidates.

3.1.3 2010 US Special Senate elections in Massachusetts and Twitter

Jessica Chung performed sentiment analysis of tweets collected from MAsen10 campaign [13]. They tried to predict election results in continuity with Tumasjan work who had claimed that share of tweets for each candidate directly corresponds to actual votes obtained by them. But this data-set brought different results for both candidates in a pre-election campaign. In order to investigate further Chung examined the tweet sentiment. They used OpinionFinder a technique to find sentiment analysis, and achieved overall accuracy of 41% which is not reliable. In order to improve accuracy they further used SentiWordNet a lexical resource with 207, 000 pair of words [13] and overall accuracy increased to 47.19%. They introduced sentiment analysis to tweets but failed to increase the accuracy with methods they adapted. Also the data-set comprised only for six days data which is somewhat insufficient.

3.1.4 Public Opinion Mining during US Presidential Elections 2008

Another study of a somewhat different kind was conducted by O'Connor in 2010 [14]. In order to automatically find public opinion, they analyzed several surveys using simple sentiment analysis methods. These surveys were on consumer confidence for presidential job approval rating for US President Barack Obama over a course of 2009 and tracking political opinions from 2008 to 2009 asking people about voting preference for Obama or McCain. They used topic based aggregation of sentiment and finally correlated sentiment obtained with opinion results for Presidential Job Approval, and with poll results for US 2008 Presidential Race. They revealed that with the application of advanced NLP techniques, costly and time consuming polls can be concretized by the text based data collected through social networks. With the inclusion of user demographics, if provided by social media sites, would strengthen the results in comparison with traditional poll results.

3.1.5 Twitter's Predictive power during US Special Senate elections 2010

Daniel Gayo-Avello used the data-set belonging to MAsen10 and USsen10 in studying predictive power of social media against several Senate races by

Figure 3.3: Evolving to Social Elections

conducting several sentiment analysis experiments [15]. According to them election results cannot be predicted using simple tweet share, sentiment analysis has to be performed for achieving better results. Daniel did not rely on studies conducted by Tumasjan and Brendan [3] [14] but also studied claims that social media cannot be used for predicting elections. On the basis of both streams, Daniel followed [14] techniques and concluded pre-election volume of tweets for MAsen10 seemed to be good for prediction of elections. They calculated MAE 17.1% following Tumasjan method i.e. Twitter volume and calculated MAE as 7.6% based on sentiment analysis. They were not satisfied with these accuracies claiming that twitter data might have been polluted by spammers and propagandists who would have tweeted intensively by creating fake accounts. Furthermore, they also discussed limits of predictability and shed light on factors and emphasized that community should have skeptical attitude towards prediction of elections results based on raw social network data.

		Recall			
classifier	accuracy	pos	neg	neu	F-score
trivial	50.19	0	0	1	0.335
MNB	62.94	0.007	0.561	0.832	0.584
ADA-MNB	65.09	0.334	0.689	0.7	0.645
SVM	64.82	0.201	0.634	0.768	0.631
ADA-SVM	64.28	0.362	0.623	0.726	0.638

Figure 3.4: Accuracy for three class sentiment classification [2]

3.1.6 Impact of Twitter for Irish General Elections 2011

The conventional Irish General Elections 2011 results were related by social analysts and researchers with combined approach involving volume based and sentiment based results obtained from twitter by Adam Bermingham and companions [2]. They collected around 32K tweets related to five main Irish parties based upon party names and abbreviations. These tweets were annotated for sentiment to relevant candidates and parties by trained annotators. To predict sentiment based on trained data they used Weka tool for classification and applied Adaboost MNB classifier which resulted in 65.09% classification accuracy as shown in Fig. 3.4[2].

3.1.7 Effectiveness of Twitter during US Midterm Elections 2010

Another detailed study for US Midterm Elections 2010 was carried out by Livne who also found cohesion in the outcomes of US midterm elections 2010 and tweets [25]. In their study, they analyzed around 460K tweets over three years for 687 candidates contesting for National House, State Governor or Senate seats. By using text and graph based techniques they tried to investigate the campaign difference between Democrats, Republicans, and other candidates and finally finding a relation among network structure, tweet contents and election results. They also proved for pair of candidates who are closer in graph distance, also have similar tweets and relate URLs contents thus forming group cohesiveness. This strong cohesiveness was more noted amongst Conservative candidates. Using the model built on the basis of graph structure, content and election results, they predicted winner or loser candidate with accuracy of around 88.0%. Their research concluded in finding out ways that Twitter based campaigns are more effective but again they did not consider content meaning, sentiment detection, which would have a

different impact on their research.

3.1.8 Comparison of French and US Presidential Elections 2012

The Comparison of 2012, French and US Presidential elections was made by Farhad [4] who analyzed the contents of tweets obtained from electoral candidates and ordinary users (voters). They adopted the approach of performing time series sentiment analysis by restricting their findings to only two candidates, Barack Obama and Mitt Romney in case for US. They also collected around 10,000 tweets for Sarkozy and Hollande and considered them as the most significant candidates for French Presidential election race. They performed time series sentiment analysis for US candidates as shown in Fig. 3.5[4] and for French candidates by computing scores, based on three difference scoring functions. These were *polarity* (counted number of positive and negative words), *sentiment* (identified positive, negative and neutral tweets), and affinity (Computes the tweets' score based on the AFINN list) *scoring functions*. During this study sudden sentiment rise in the tweets for French election campaign was also noticed, upon investigating they discovered that this happened due to a speculation by a candidate Eva Joly, which resulted in positive opinion for Sarkozy and less impact for Hollande. They also evaluated effect of allegation on twitter stream, made on Sarkozy for receiving 42M pounds from Gaddafi for his election campaign in 2007. This allegation played decisive role for Hollande's victory, this was reflected as a negative sentiment for Sarkozy. Besides sentiment analysis they also performed word-cloud analysis in order to find popular topics being discussed in tweets. They found most frequent terms used by people in their tweets by 'BOW' approach and were as President, Vote, America, Job, Tax etc. for US elections and for French elections the tweets revolved around topics such as Gaddafi, Receive, Funding, and claim etc. In conclusion they claimed twitter to be a best medium for candidates in judging and responding their voters and vice versa but once again they considered selected tweets for selected candidates by ignoring all others, also they ignored neutral tweets while performing time series sentiment analysis.

3.2 Text Classifications

Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a free document to a set of predefined class. The

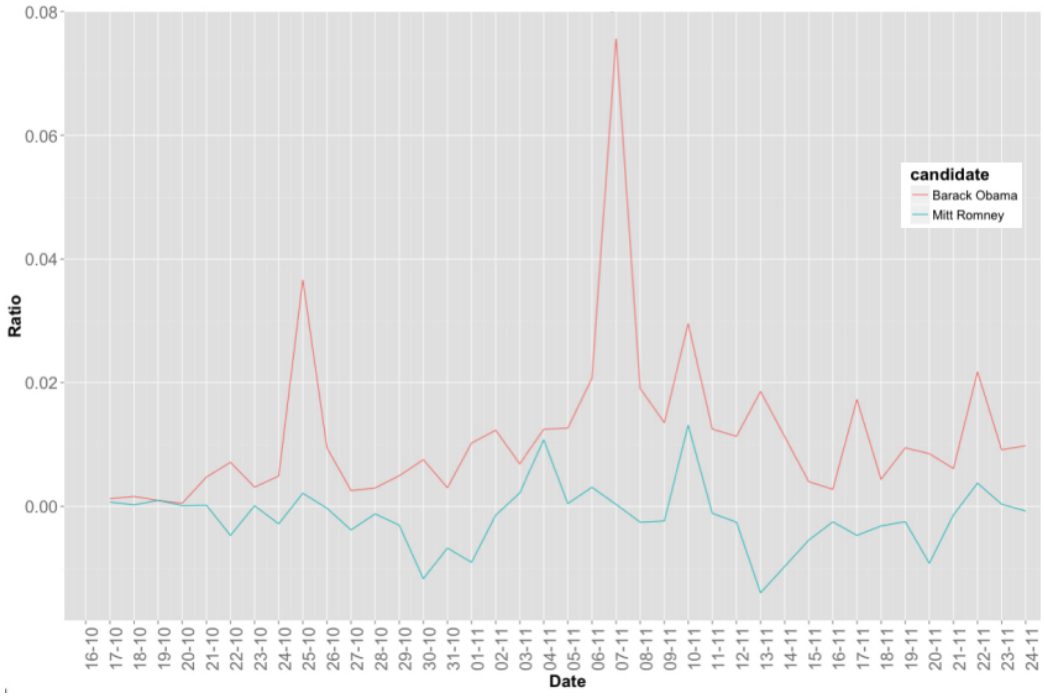

Figure 3.5: Sentiments analysis for Obama and Romney [4]

class is determined based on its textual content and extracted features. TC plays vital role in content management, sentiment mining and analysis, reviews analysis, spam filtering, categorization of news articles, organization of web pages and with other various applications.

3.3 Statistical Text Classification

In statistical Text classification, machine learning methods are applied on free documents for categorization. These machine learning methods use classification rules based on human labeled documents.

3.3.1 Binary Classification

Having only two classes in text classification task is called binary classification. The predicted class for a test document would either fall in class A or Class B. Examples for Binary classification would be like Spam/non-spam, has / has not, True/False, Positive/Negative etc.

3.3.2 Multi-class Classification

Having more than two classes in all and classifying test documents for one of them is called multi-class classification. Examples of multi-class would be like positive/negative/neutral etc.

3.4 Generic Strategy for Text Classification

Main steps involved in generic text classification are as follows:-

3.4.1 Pre-processing

Raw text contents from any source contain lots of words / characters which are not beneficial to text classification, thus have to be removed. The size of input data decreases considerably by using data pre-processing techniques. [24]

- It involves removal of stopwords as they are not useful for classification. (for example, 'a', 'an', 'the', 'of', 'for', 'if', 'I', etc.)

- Stemming is a process of removal for derivational affixes from words, through which words are converted to its base word. Some of the famous stemmer are Lovins Stemmer, Porter Stemmer, Paice Stemmer etc. [31] [`http://informationretrieval.org/`]
  (for example Porter Stemmer would stem 'generalizations' ⇒ 'generalization' ⇒ 'generalize' ⇒ 'general' ⇒ 'gener')

- Additionally pre-processing can involve removal of extra whitespaces, punctuation, repeated words etc.

3.4.2 Feature Formation, Selection / Extraction

In machine learning, classified atomic item is called 'feature vector'. A simple way to convert a document into feature vector is 'Bag-of-words' technique. The document act as string of tokens regardless of word order and grammar.

Identification of important tokens and limit the dimensionality is done through feature selection. This uses various weighting methods like TF-IDF (term frequency-inverse document frequency), LSI (latent semantic indexing), multi-word [24]. Feature reductions are performed using χ^2, PCA (Principal components analysis), information gain, mutual information etc.

3.5 Type of Classifiers

In literature various supervised methods are proposed for machine learning in classifying text documents. Probabilistic text classification is provided by Naïve Bayesian classifier which is based upon conditional independence among feature vectors. NB classifier operates on categories having sufficient training instances for each category. Support Vector Machines provide efficient results for binary classification but not reliable for mutli-classification. Some of the other classifiers are decision trees using C4.5 and ID3. Decision trees perform well for fewer features but becomes difficult in managing large number of features. KNN algorithm assigns most frequent class amongst k nearest training examples to unseen example.

3.5.1 Naïve Bayes Text Classification

Naïve Bayes text classification is based on Bayes Theorem, this can be explained using 3.1. Let T be considered as evidence, and C considered as hypothesis such that data tuple D (tweet) belongs to specified class C. P(C|T) is called **posterior probability** of C conditioned on T, where P(T) and P(C) are prior probabilities of T, T respectively. [8]

$$P(C|T) = \frac{P(C)P(T|C)}{P(T)} \tag{3.1}$$

P(T|C) represents the probability of tweet given class and tweets can be modeled as set of words, thus P(T|C) can be written as:

$$P(T|C) = \prod_i P(W_i|C) \tag{3.2}$$

So

$$P(C|T) = P(C)\prod_i P(W_i|C) \tag{3.3}$$

Where P(W_i|C) is the probability that the i^{th} word of a given tweet occurs in a tweet from class C. This approach is efficient with less computational time for training and easy to construct and can be applied to large data set.

3.5.2 Naïve Bayes Multinominal

NBMN has been for TC because of its efficiency and simplicity, by capturing word frequency information from documents. NBMN achieves text classification using NB learning but with additional specific parametric model learned from training documents as shown in 3.4, here f_i is frequency of word w_i in a document 'd'. P(w_i|c) is the conditional probability that a word w_i may happen in a document d given the class value c, and n is the number of unique words appearing in the document.

$$P(C|D) = \frac{P(C)\prod_{i=1}^{n} P(W_i|C)^{f_i}}{P(d)} \tag{3.4}$$

3.5.3 Random Forest

Random forest (or random forests) is an ensemble learning based classifier consisting of many decision tree models (Breiman, 2001). Ensemble learning is based on two methods 'boosting' with successive trees dependent on earlier trees and 'bagging' in which successive trees are not dependent on earlier trees. RF are good example of ensemble method which combines predictions of weak classifiers. RF combines bagging method with random selection of features from all chosen features set $\{A_i\}_{i=1}^{n}$. At each node as in Fig. 3.6 RF makes two decisions, one of them is selection of number of best features in the random subset and the number of trees in the forest. The terminal nodes are assigned the class [10].

3.5.4 k Nearest Neighbor

KNN is non-parametric, instance based lazy learners. In a Nearest neighbor classifier, a test tuple is compared with training tuples that are similar to it. The test tuple(s) and training tuple(s) are represented as n attributes / feature vectors. These n attributes are stored in an n-dimensional pattern space. Let T be a test tuple, in order to find class for test tuple T, KNN searches for closest k training tuples represented as n-dimensional pattern space. The closeness of k training tuples w.r.t. test tuple T is measured in terms of Euclidean distance as mentioned below in equation 3.1. Let X_1 =

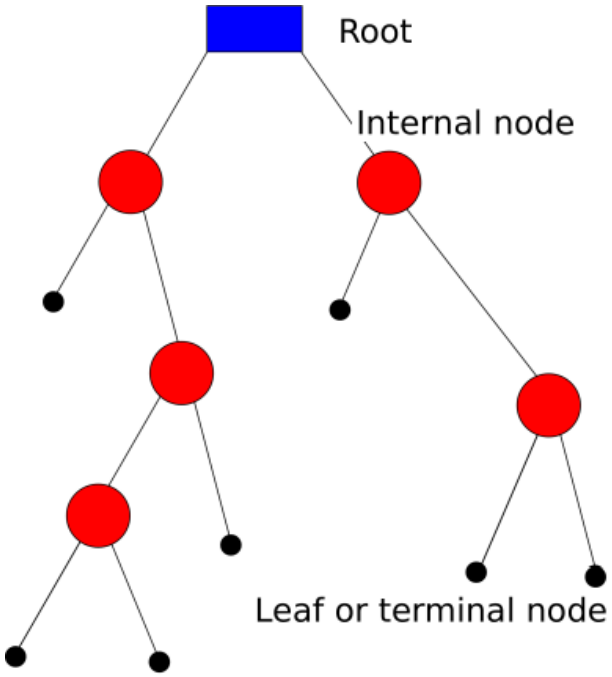

Figure 3.6: Decision Tree Structure [10]

(x_{11}, x_{12}, ... , x_{1n}) and X_2 = (x_{21}, x_{22}, ... , x_{2n}) be tuples under consideration with n attributes, their distance can be measured as [10] [26]:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{3.5}$$

3.5.5 Support Vector Machines

SVM is another supervised machine learning classifier used to analyze, model and classify the data. The idea for SVM was first given by Vapnik and Lerner in 1963, in which a hyper-plane is created between two sets of data, identified by two classes. While separating linear data there are infinite ways to construct hyper-planes as shown in Fig. 3.7 with 4 hyper-planes drawn separating two classes.

Optimal hyper-planes are constructed in way that space between point closes to the hyper-plane and the hyper-plane itself is maximized [26], this distance so obtained is called optimal margin as shown in Fig. 3.8. The vectors touching the maximized margins are called support vectors. These support vector are difficult to classify but rich in classification information.

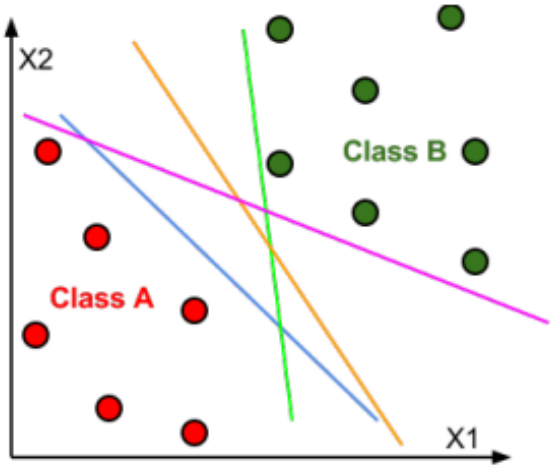

Figure 3.7: Linearly separable data with 4 hyper-planes

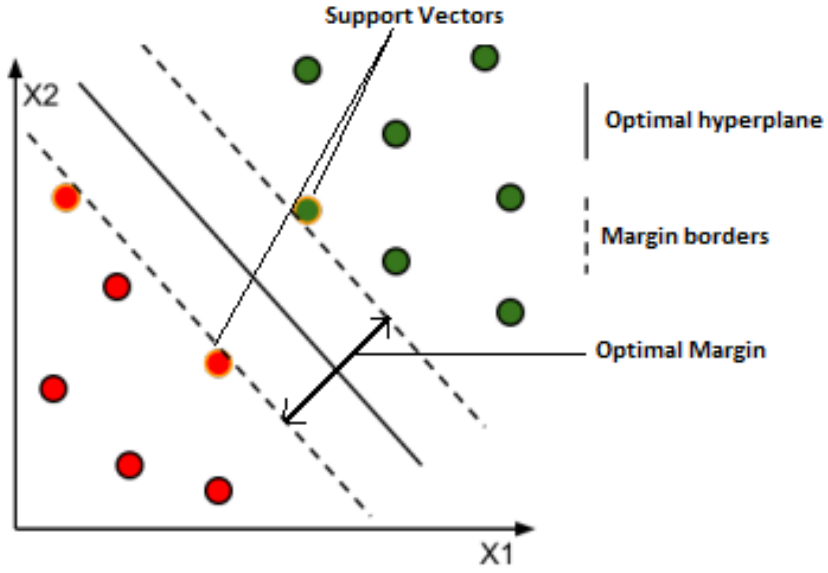

Figure 3.8: Optimal placement of hyper-planes [26]

A maximized margin would increase chances for correct classification of new vector to respective class even in the presence of noise.

SVM converts all data points in m feature vectors. If training data is linearly separable then a pair (w,b) exists with 'w' as weight and 'b' as threshold, [8][26]

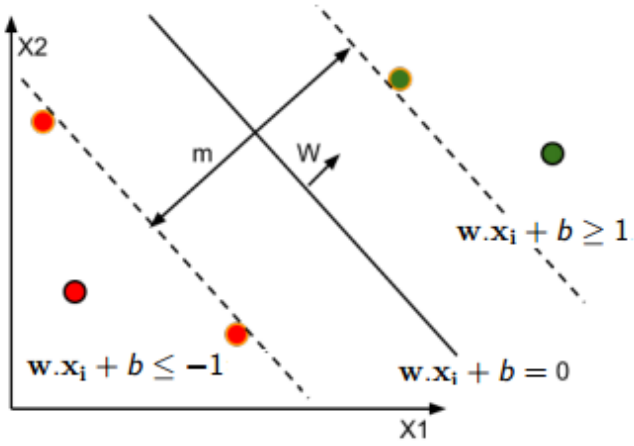

Figure 3.9: Support Vectors touching maximized margins [26]

$$\mathbf{w}.\mathbf{x_i} + b \geq 1 \text{ for } y_i = 1 \tag{3.6}$$

$$\mathbf{w}.\mathbf{x_i} + b \leq -1 \text{ for } y_i = -1 \tag{3.7}$$

When considering multi-class classification, SVM consider pairs from all classes and divides training vectors into classes separated by maximum margins. Due to this computationally SVM becomes expensive as it involves large amount of training time.

3.6 Short Text Classification

Traditional techniques as discussed in previous sections 'Bag of Words' performs well with large documents which are rich in content and extraction of feature vector is much easier whereas, small text documents like SMS, Tweet text, forum posts etc. contains very little information, thus classification of small text document itself a big challenge.

3.7 Performance Measures

Text classification is evaluated using performance measures. Common metrics involving text classification includes recall, precision, accuracy, error rate and F measure.

Table 3.1: Confusion Matrix.

		Actual Class		
		Positive	Negative	Total
Predicted Class	**Positive**	TP	FP	PP
	Negative	FN	TN	NP
	Total	P	N	

3.7.1 Precision and Recall

Precision, recall and accuracy are most effective method among various methods used in estimating effectiveness of text classification. In order to calculate for these methods we must have an idea for true positive (TP) which means document classified correctly w.r.t. to category. False positive means document classified incorrectly relating to category. Similarly False negative (FN) and true negative (TN) means determined document is not assigned category but should be and determined document should not be marked as being in a particular category and are not respectively. Confusion matrix is given in table 4.1.

The exactness of classifier is measured through precision as given in equation 3.8[26], which clearly shows greater the TP greater would be precision whereas greater FP would lower the precision. So in actual precision means 'what percent of positive predictions were correct?'.

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

Also Recall measures completeness of classifier. Lower number of FN would guarantee higher classifier recall. Recall would be defined as 'what percent of positive cases were caught?'

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

Accuracy estimates 'what percent of predictions were correct'.

$$Accuracy = \frac{TP + TN}{N} \tag{3.10}$$

3.7.2 Combination of Precision and Recall

Two metrics as in equations 3.8-3.9, when combined together as in equation 3.11 makes important metric called F measure / F-score, which is weighted harmonic mean of precision and recall. F score is measure of test accuracy and is used to evaluate text classification system [8][26].

$$Fmeasure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3.11}$$

Chapter 4

Methodology

4.1 Twitter Corpus Collection

To study qualitative behavior of twitter by using twitter API we have collected 612,802 tweets with average tweets per day mentions in Fig. 4.1. These tweets were based on switches / keywords. These keywords comprised of full names and acronyms of different political parties on national level for Pakistan, like Pakistan People's Party (Parliamentarian) or PPP(P), Pakistan Muslim League (Nawaz) or PML(N), Pakistan Tehreek-e-Insaf or PTI etc. The keywords also contained names of party leaders for major political parties of Pakistan on National Level and Provisional Level, like Nawaz Sharif, Imran Khan, or Benazir Bhutto etc. These names were taken from election commission of Pakistan's website. In addition to party and leadership names, names of provinces, major cities and party symbols were chosen as keywords as in Table. 4.1. This resulted in downloading huge amount of relevant and irrelevant tweets based upon keywords. Here irrelevant means same keywords but from other geographical region, and in different languages. The regions were verified from the locations provided by the twitter users in their profile. These locations were cleaned for misspells and annotated before storing in repository in order to identify the geography of the tweet.

4.2 Pre-processing Tasks

The tweets data downloaded were in JavaScript Object Notation (JSON) format shown in Fig. 4.2. We identified relevant fields for extraction which could be useful during classification and further analysis of tweets. These included, demographics related to tweets such, tweet identification (id), text,

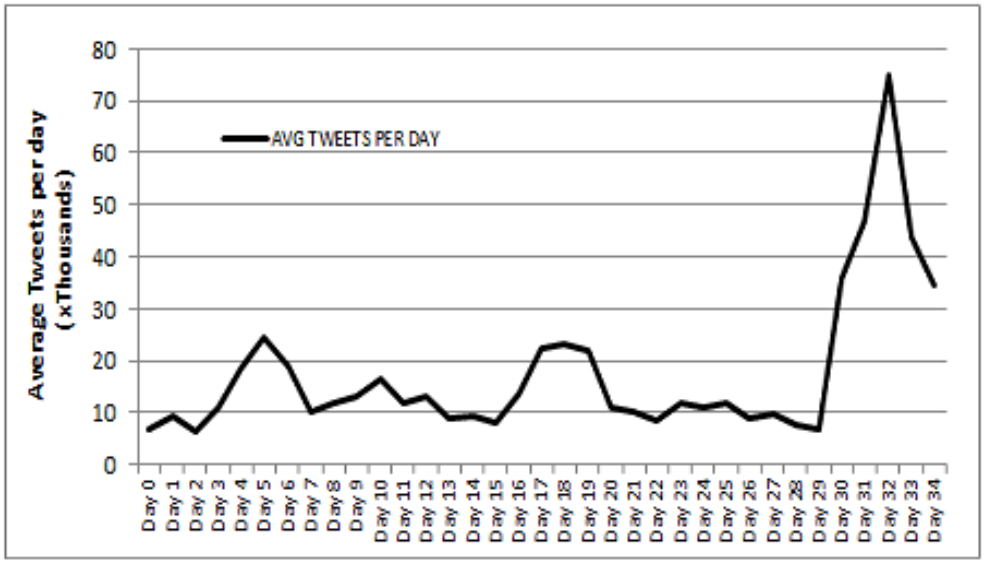

Figure 4.1: Average tweets per day

creation time etc., The fields related to user, contains details of twitter, user identification (id), user creation time, status count, followers count, following count etc. Twitter API entities' part was used to extract fields like hashtags, mentions, URLs, etc, for our further research. Finally user's location information was extracted from JSON part as shown in Fig. 4.3. For Classification purpose we required only Text part of JSON format whereas rest of information was kept in repository for further analysis shown in Fig. 4.4.

4.3 Tweets Annotations

Previous studies suggested that emoticons are true reflector for sentiment but in some cases emoticons can lead to sentiment misclassification [22]. In order to overcome sentiment inconsistency with emoticons we considered manual labeling of the tweet text data. The tweets projecting party manifesto in a positive way, containing motivations to vote for a particular party or leadership having party name or leadership in a tweet text. Tweets projecting either appreciation or showing satisfaction with for some party were labeled as positive tweet. In contrary to this the tweets containing negative words, emoticons, for certain party were labeled as negative tweet for that particular party. The tweets text showed dissatisfaction with certain party, its agenda

Table 4.1: Examples of Keywords used for Tweet Selection

PML	PTI	Election 2013
Zardari	Benazir Bhutto	MQM
Pakistan Peoples Party	Awami National Party	Mutahida Qaumi Movement
Pakistan Tehreek-e-Insaf	Jama'at-e-Islami	Jamiat Ulema-e-Islam
JUI	All Pakistan Muslim League	Bhutto
Punjab	Sindh	Sind
Elections in Pakistan	Islamabad	APML
Karachi	Peshawar	Lahore
PML	PPP(P)	JI
PPP	PML(N)	Pak Election
Nawaz Sharif	Chaudary Brothers	Sharif Brothers
Imran khan	Shabaz Sharif	KPK
Baluchistan	ANP	Pakistan
Quetta	Lion	Teer


u'created_at': u'Fri Jun 27 18:03:32 +0000 2008', u'contributors_enabled': True, u'time_zone': u'Eastern Time (US & Canada)',
u'profile_sidebar_border_color': u'FFFFFF', u'default_profile': False, u'following': None, u'listed_count': 1406}, u'geo':
None, u'in_reply_to_user_id_str': None, u'possibly_sensitive': False, u'created_at': u'Mon Oct 15 14:10:05 +0000 2012',
u'possibly_sensitive_editable': True, u'in_reply_to_status_id_str': None, u'place': None}, u'user': {u'follow_request_sent':
None, u'profile_use_background_image': True, u'id': 116517141, u'verified': False, u'profile_image_url_https':
u'https://si0.twimg.com/profile_images/2328184850/2M2TGAmq_normal', u'profile_sidebar_fill_color': u'D60AFF',
u'is_translator': False, u'profile_text_color': u'0A0101', u'followers_count': 297, u'protected': False, u'id_str':
u'116517141', u'default_profile_image': False, u'location': u'', u'utc_offset': -21600, u'statuses_count': 15942,
u'description': u'Music fan, writer, grad student, avid reader, mom and huge Preds fan.', u'friends_count': 705,
u'profile_background_image_url_https': u'https://si0.twimg.com/profile_background_images/99089675/st._john.jpg',
u'profile_link_color': u'05CDFF', u'profile_image_url': u'http://a0.twimg.com/profile_images/2328184850/2M2TGAmq_normal',
u'notifications': None, u'geo_enabled': False, u'profile_background_color': u'591CFF', u'profile_background_image_url':
u'http://a0.twimg.com/profile_background_images/99089675/st._john.jpg', u'name': u'Dawn', u'lang': u'en',
u'profile_background_tile': True, u'favourites_count': 119, u'screen_name': u'jab0217', u'url':
u'http://dawnshiver.tumblr.com', u'created_at': u'Mon Feb 22 19:02:41 +0000 2010', u'contributors_enabled': False,
u'time_zone': u'Central Time (US & Canada)', u'profile_sidebar_border_color': u'0335FF', u'default_profile': False,
u'following': None, u'listed_count': 15}, u'geo': None, u'in_reply_to_user_id_str': None, u'possibly_sensitive': False,
u'created_at': u'Mon Oct 15 14:31:14 +0000 2012', u'possibly_sensitive_editable': True, u'in_reply_to_status_id_str': None,
u'place': None}
{u'user': {u'follow_request_sent': None, u'profile_use_background_image': True, u'id': 274290600, u'verified': False,
u'profile_image_url_https': u'https://si0.twimg.com/profile_images/2685195316/061c90de07003acb28dc6cb216caa133_normal.jpeg',
u'profile_sidebar_fill_color': u'DDEEF6', u'is_translator': False, u'profile_text_color': u'333333', u'followers_count':
2549, u'protected': False, u'id_str': u'274290600', u'default_profile_image': False, u'location': u'DTH , East Orange',
u'utc_offset': None, u'statuses_count': 27385, u'description': u'reguardless of how it goes down life goes on \r\nremember
that and you will succeed', u'friends_count': 1571, u'profile_background_image_url_https':
u'https://si0.twimg.com/profile_background_images/627338084/90ubrcl5hccq03xr6iig.jpeg', u'profile_link_color': u'0084B4',
u'profile_image_url': u'http://a0.twimg.com/profile_images/2685195316/061c90de07003acb28dc6cb216caa133_normal.jpeg',
u'notifications': None, u'geo_enabled': False, u'profile_background_color': u'C0DEED', u'profile_background_image_url':
u'http://a0.twimg.com/profile_background_images/627338084/90ubrcl5hccq03xr6iig.jpeg', u'name': u'William E. Kornegay',
u'lang': u'en', u'profile_background_tile': True, u'favourites_count': 0, u'screen_name': u'Polo_Smash', u'url': None,
u'created_at': u'Wed Mar 30 02:37:59 +0000 2011', u'contributors_enabled': False, u'time_zone': None,
u'profile_sidebar_border_color': u'C0DEED', u'default_profile': False, u'following': None, u'listed_count': 0}, u'favorited':

Figure 4.2: Data obtained in JSON format through API

or leadership over past performance. Similarly tweets which did not show any tilt towards any party but it was related to general elections campaign were termed as neutral tweets. In this way we got politically positive, negative and neutral sentiments tweets for several parties in our data. Some of the

Fri Mar 01 14:41:01 +0000 2013	182405906	False	RT @sigmanewsdotus: APC JUI-F Serukan Perundingan Damai dengan Tehreek-e-Taliban Pakistan - See more at:... http://t.co/FHjg1qB0hv
Fri Mar 01 14:41:01 +0000 2013	122089147	False	Woke up to find out Musharraf is coming home, Spring is just around the corner & ofcourse ECP suggests PTI is disqualified. Haha ya right!
Fri Mar 01 14:41:04 +0000 2013	145926781	False	RT @BabyBhutto: Election stress causing hair loss! Uncle Nawaz contacts Sahir Lodhi for advice ..... #Pakistan #pmln #pti Nawaz Sharif # ...
Fri Mar 01 14:41:07 +0000 2013	320713565	False	PTI is not qualified yet to participate in Gen.Elect. Election Commission of Pakistan
Fri Mar 01 14:41:07 +0000 2013	428813200	False	RT @ImranInc: After #pti Intra Party Elections, #pmln is also conducting their Intra Party Selection in Raiwind... #Pakistan #PTIFamily ...
Fri Mar 01 14:41:08 +0000 2013	148246711	False	RT @saqibtaj: By the grace of Allah almighty our panel has clean sweeped Narowal District in PTI intra party election. I have... http:// ...
Fri Mar 01 14:41:08 +0000 2013	182405906	False	RT @sigmanewsdotus: JUI-Fs APC calls for immediate peace talks with TTP - See more at: http://t.co/Zd1E3AfmFD
Fri Mar 01 14:41:09 +0000 2013	408942066	False	@HilmiD ehhh kau>< muke aku mmg macam diyer pe:ppp heh ~~î î î î~~ main main aje:p
Fri Mar 01 14:41:11 +0000 2013	461277652	False	Ijlaas Sec Mem & Unit Mem at Baghe Mustafa Ltaifabad No 8 HYDERABAD ZONE MQM http://t.co/ApPywSFsJF
Fri Mar 01 14:41:12 +0000 2013	864920743	False	RT @allaboutmqm: Social Media Camp in 26th Labour Convention of MQM Labour Division tweeting for #MQMLabourRights http://t.co/bWl6Wd3pSe ...
Fri Mar 01 14:41:12 +0000 2013	945929700	False	we cannot understand MQM even after understnding it
Fri Mar 01 14:41:14 +0000 2013	917002142	False	#MQM walks out of NA to oppose petrol price hike http://t.co/eMBaB1FzYE â€¦ #Karachi #Pakistan #MQMLabourRights
Fri Mar 01 14:41:14 +0000 2013	60856622	False	RT @BabyBhutto: Yo Stupid Karachites! Your lives are worth less than the paper my Daddy wrote my Mummy's will on .. go pay bhatta! #MQM ...
Fri Mar 01 14:41:16 +0000 2013	1179053299	False	@CMShehbaz to permenant 100k employees near elections is not a justifiable decision, as these type of norms are of PPP.

Figure 4.3: Relevant fields extracted from JSON format

examples are shown in Table 5.1. We started with labeling of approx. 3600 tweets with 1200 tweets (approx. 33.33%) for each positive, negative and neutral class. These tweets were labeled by three different persons and finally their results were combined to set a basis for training data. The labeled data in a text file with each line representing individual tweet, was then pre-processed with the removal of following, steps also shown in Fig. 4.5.

- Duplicate tweets (If any),
- URLs, mentions words starting with '@' sign (Kept separately for further analysis),
- Hashtags, words starting with '#' symbol (Kept separately for further analysis)
- Extra whitespaces,
- Repeated words, words starting with number,
- Small words (threshold set 3,)
- Punctuations (?.,)

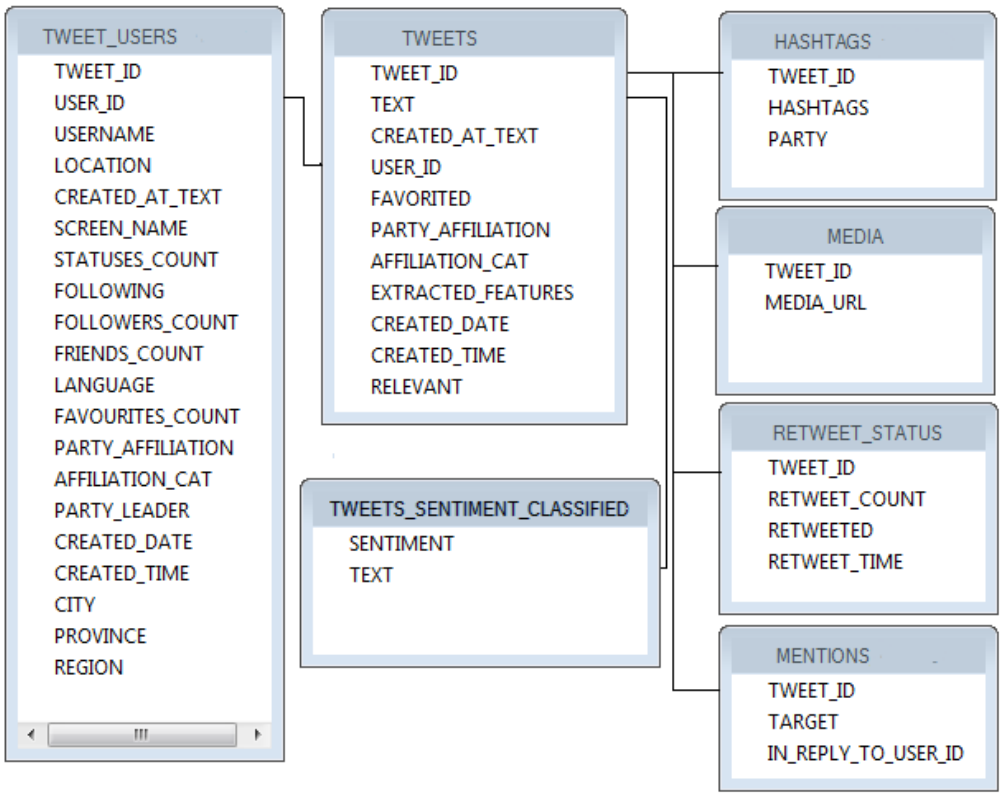

Figure 4.4: Schema to store extracted fields

Finally whole tweet text data was converted to lower case.

Data pre-processing step reduced the size of text considerably. Now we have three folders based on our three preset classes positive, negative and neutral. The cleaned labeled dataset is bifurcated with each tweet to be stored in separate text document using python. These tweets are placed in their respective folders according to their label/class as in Fig. 4.5. Each of These three folders sentiment positive, negative and neutral contained around 1200 tweets text documents which were later used by Rainbow program and Weka tool for classification as a training purpose. A brief for both tools is under.

Table 4.2: Sample Tweets Labeled Pos, Neg and Neu

Sentiment	Tweets
—Positive—	—I am not a rootless phenomenon. I am not going to run away from my country. I am not leaving my roots Zulfikar Ali Bhutto #VoteTeerKa—
—Positive—	—#Altaf Hussain Bhai is a Great Full Leader in the World #MQM is a Peaceful Party In The #Pakistan Vote 4 #Kite #MQM #Pakistan—
—Positive—	—- Imran Khan.. you're the bestt..—
—Positive—	—#Change #PTI #Imrankhan #PakVotes #Elections2013 http://t.co/rBZZRcDDuz—
—Positive—	—#Election2013 Time to rethink your strategy #MQM the only solution @wasimz http://t.co/aZhZLizWN8—
—Negative—	—#ANP & elders of society decide not to let women vote in upcoming election in Matwani Village—
—Negative—	—#KarachiShutdown had a call from a friend today he wants to move to Punjab ... Sad because I always wanted to move to karachi—
—Negative—	—#MQM called for strike and protests #irony—
—Negative—	—#MQM to challenge #Karachi delimitation:#AltafHussain http://t.co/JIFk1gxfYS #NoToDelimitation #Pakistan—
—Negative—	—#Pakistan #Election ECP seeks utility bill defaulters s lists: Image: http://t.co/Zhno1u72oZ... http://t.co/LrV4xngDZS #PTI #PPP #PMLN—
—Neutral—	—#pakistan People of Rawalpindi Will Vote For??? http://t.co/LpoJqxZ2Sa #ppp #pmln #pti #mqm—
—Neutral—	—#PPP has constituted a committee for seat adjustment with different parties in #Balochistan—
—Neutral—	—#TabdeeliRazakar Imran Khan Address in launching ceremony of Tabdeeli Razakar - 30th March 2013 http://t.co/wAKGQkPYGr—
—Neutral—	—: Altaf Hussain appeals people to donate for election expenses #MQM #vote4patang—
—Neutral—	—@BeingUnknownPTI @SalooDurrani Did you check @mrsultan713 analysis of PPP/PML economic policies?:http://t.co/S9KaTT1CLq—

4.4 Tools Used

Following Rainbow and Weka tools were selected for training and testing data-set obtained with different classifiers.

Rainbow: Rainbow based on 'Bag of words' (BOW) technique is program used for text classification. Rainbow reads all document corpus and converts it to model containing statistics pertaining to each of labeled document. Classifier is run on the basis of the obtained model with train and test

data, which results in accuracies for correct versus incorrect classification.

Weka: Based on Java, is data mining tool equipped with numerous machine learning algorithms, which can be directly applied to datasets. It has the provision of performing all tasks related to pre-processing, classification, clustering etc.

4.5 Algorithms & Approaches Used

Using Rainbow we used classifications methods NB, KNN, and Prind. Each classifier was tested for its accuracy for different preset parameters. We took set of words (unigram) with highest mutual information as features vectors ranging from 10 to 200. In order to smooth word probabilities we used 'Laplace' method which helped in avoiding zero values. Stemming was performed with 'Porter Stemmer'. The accuracies were obtained for positive versus negative sentiments, positive, negative and neutral sentiments separately with 40% data placed in test set and remaining for training set. As in Fig. 4.6, we can see that an average accuracy of 70% and above was obtained for two classes positive and negative. With the inclusion of neutral class our accuracies dropped because we have tweets which are neither positive nor negative. Their importance cannot be ignored thus extending our problem from binary text classification (i.e. Positive / Negative) to multi-class problem (i.e. Positive / Negative / Neutral).

Same data set was tested with some more algorithms like RF, SVM, NB and NBMN for supervised machine learning classification. We calculated classification accuracies using different set of parameters, TF-IDF transformations for finding the words that were strongly related to relevant documents. The reduced feature vectors after stemming with set of attributes ranging from 10 to 200 were given to classifiers for prediction purpose. We used 5-fold cross validation in case of Weka. Fig. 4.7 shows increase in accuracy with increase in number of attributes. NB performed best with few more percent on average accuracy for both sets of sentiments from rest of classifiers. RF & SVM were almost together with accuracies. We used RF a decision tree learner, different tree (10 trees in our case) were built with each tree predicting a number using random samples and random features. For overall prediction an average predictions is obtained from all individual tree.

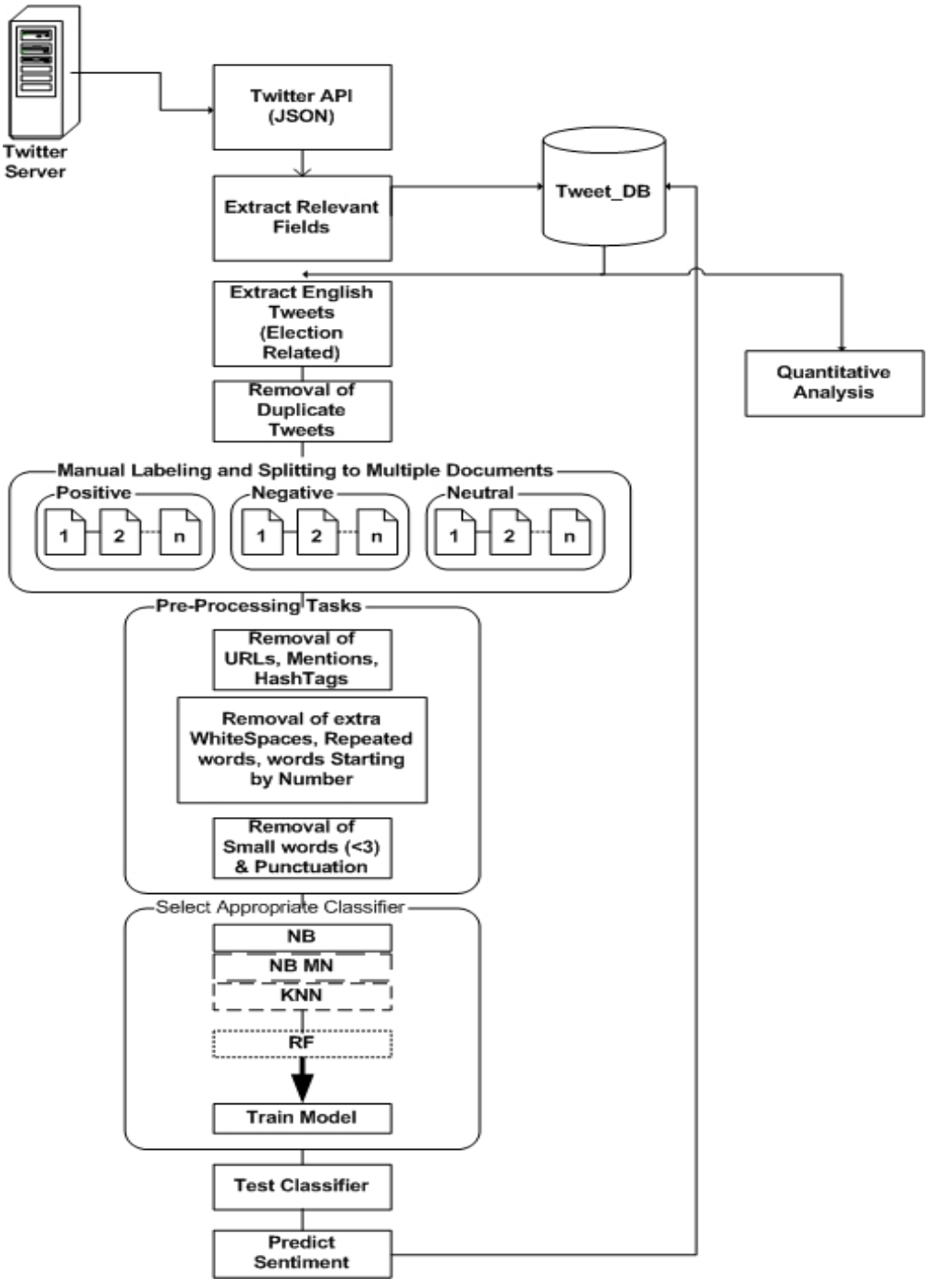

Figure 4.5: Architecture for Sentiment Prediction

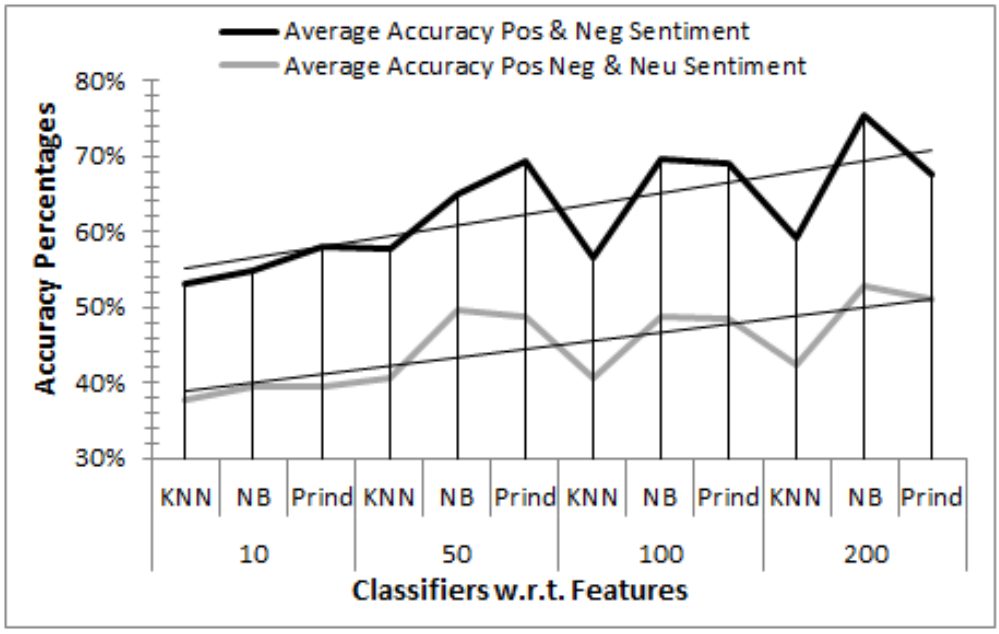

Figure 4.6: Effect of Feature Size for Rainbow

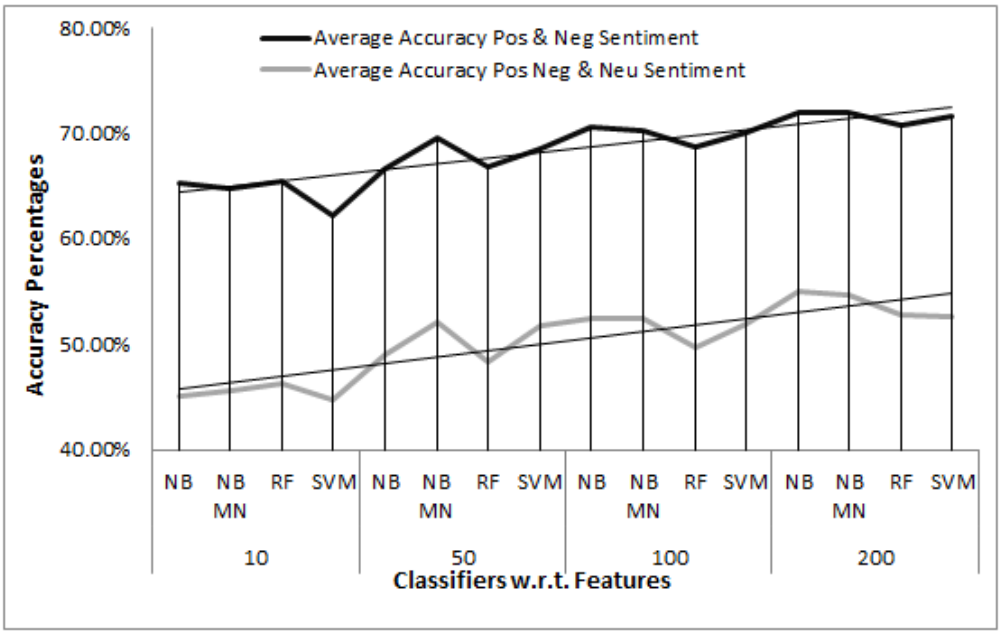

Figure 4.7: Effect of Feature Size for Weka

4.6 Study of Quantitative Behavior

Study of quantitative behavior for twitter corpus involved in predicting sentiment for whole relevant tweets with extensive data cleaning, labeling for

other relevant fields for better analysis purpose. This relevant tagging of tweets was based on English language, region, country, as we were more concerned with tweets related to Pakistan's General Elections. We chose NB Classifier as it showed better average accuracies both for Rainbow and Weka to classify our tweets based on already trained labeled data. After sentiment prediction using NB classifier and thorough annotations we were left with 226,510 politically positive, negative and neutral tweets (out of total tweets collected through API).

User participation can be easily identified by number of political tweets, as it contained mentions i.e. words starting with @sign. These mentions contained political parties name, political celebrities and people involved in political campaign. The cleaned data contained about 42,270 tweets started with mention, related to Pakistan's political scenario. This showed direct involvement of people who not only share their views but also involved others in this campaign. About 18,225 tweets contained mention in the tweet text. This showed that tweets sent to different individuals which lead towards group discussion regarding any political development. There are tweets which users sent as it is, to their followers called re-tweets. In re-tweets without making any addition the user forward received tweet to their followers by giving weightage to one's opinion. This could be text, a URL linking other website, media or image. In our relevant tweet data about 50, 354 tweets were found re-tweets. There were 37, 564 re-tweets with simple text, 12,790 tweets containing URLs. This helped us in finding out that simple text was more re-tweeted than tweets containing hyperlinks. Next we saw whether these tweets were biased, or there was some political propaganda, we have to identify the users, whether if there are couple of users who are tweeting and politicizing the forum, or it is an open discussion.

Chapter 5

Experimental Results

In this chapter, the experimental results for techniques described in previous chapters are presented. The data sets used in these experiments are tweets along with different attributes obtained from Twitter. Considering relevant political tweets for Pakistan Election 2013, classified sentiment for each tweet using NB Classifier. This sentiment prediction became basis for our further discussion.

5.1 Predicting Election Outcomes

In subsequent subsections province wise sentiment classified for different parties are discussed. Also how twitter users participated for provincial capitals was discussed with the help of graphs.

5.1.1 Province wise Sentiment prediction for Parties

By keeping identity of each tweet in contact we related user information, political party information (extracted from hashtags used, party names, mentions) with predicted sentiment. As we already annotated our data province wise, and city wise, so this helped us in drilling down from National Level to Province Level, from political party to individual twitter user belonging to them. By having in depth analysis, we came to know there are majorly two streams of users participating on twitter discussions, one of them with locations not related to Pakistanis cities, they were termed as 'Foreign Pakistanis' living abroad. They could only participate in discussions on social media as no law provision existed in 2013 for Pakistani's living abroad to cast their votes. It became clear that twitter users participating from outside Pakistan

participated in a positive manner. The other stream belongs to 'local users' within domain of Pakistan, from all four provinces and different cities. In the Fig. 5.1, we can find positive participation of foreign users sentiment wise for each political party. The Fig. 5.1 also shows sentiment detection for local users belonging to Federal Capital and other provinces. These graphs showed contribution of positive discussion majority from Independent (General public with no political party affiliation) PTI, MQM supporters mostly on the top for every province and capital. These parties campaigned positively on twitter by setting accounts for party and party leader ship. We detected positive attitude of Pakistani people towards Pre-election campaign as an answer to first research question. On the other side, no significant share was contributed from religious political parties from any of the province. The reason behind could be twitter users did not show interest in their manifesto or these parties did not actively participated over twitter for their campaign. By using our results we have seen an interesting fact regarding political party Pakistan Awami Tehreek' which boycotted Elections 2013 but its root causes for not participating were taken with positive sentiment. We noticed less number of twitter participants for province Baluchistan and KPK, the reason behind could be low population and literacy rate.

5.1.2 Twitter User Participation for Capital and Provisional Capitals

We have found as mentioned in Fig. 5.2, that twitter users actively participated over its platform for election campaign from all part of Pakistan. We can see the party-wise trends for Federal Capital and Provisional Capitals. If we consider Islamabad, we can see Independent twitter users and PTI supporters are on the top with MQM and PPP(P) side by side. For provisional capital Karachi, besides Independent participants, PTI supporters are head to head with MQM for twitter. For Punjab's provincial capital Lahore, we can once again see Independent participants on the top with PTI and PMLN supporter in a handsome number, same could be seen for KPK and Baluchistan Province.

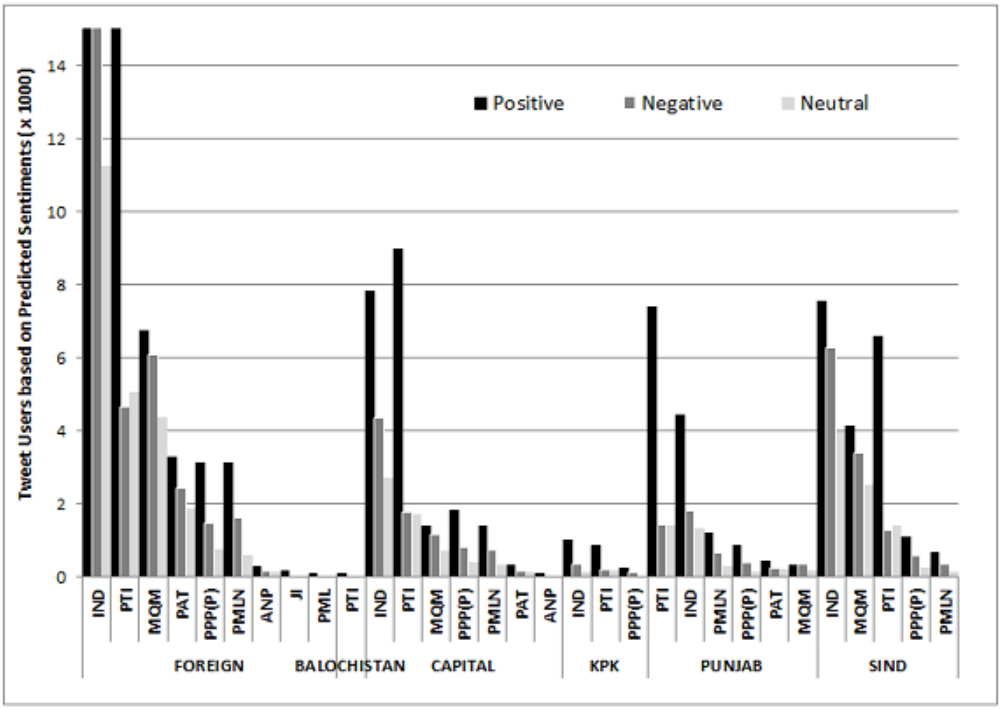

Figure 5.1: Province wise Sentiment prediction for Parties.

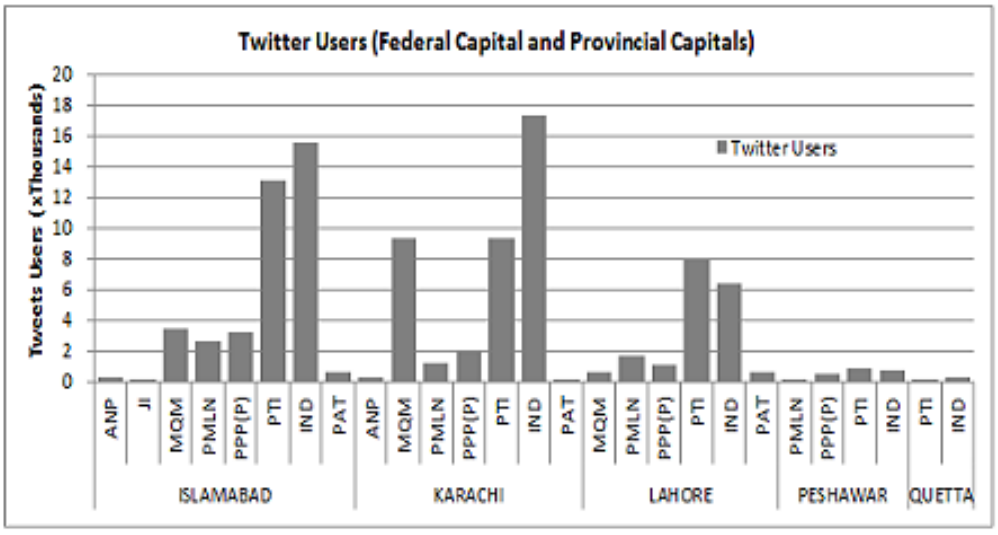

Figure 5.2: Twitter User Participation Capital and Provisional Capitals.

5.2 Comparative results for ECP and Twitter Findings

A comparison of actual ECP results and sentiment obtained from our results for top 5 parties is discussed in following subsections. Also relation a relation

was established among actual polled votes and twitter users who participated in social campaign for their political parties.

5.2.1 Top 5 Parties with Sentiment Tweets and actual Polled votes

Election Commission of Pakistan (ECP) after successful holding the Elections for Pakistan in May 2013, published complete results [29]. ECP published Top political parties securing votes from all over Pakistan, the percentages are shown in Table 6.1. This shows sentiment-wise participation percentages along with percentages of actual polled votes. We can relate positive sentiment with actual votes gained by the political party with some differences against PTI and Independent twitter users and voters. This shows PTI had more twitter users over social media campaigning positively. PTI went positively over the twitter after the incident of its leader Imran Khan' falling from Lift in a campaign gathering. This event was taken positively by twitter users resulting as a spike as in Fig. 4.1, and increased twitter user participation as shown in Fig. 5.3. This also gives answer to research question 3.

Table 5.1: Top 5 Parties with Sentiment Tweets and actual Polled votes.

Party	**Pos**	**Neg**	**Neu**	**Actual Polled Votes**
PTI	72.29%	13.57%	14.14%	20.32%
PPP(P)	59.38%	27.80%	12.82%	18.28%
PMLN	57.50%	30.03%	12.47%	39.35%
IND	49.72%	30.44%	19.83%	15.56%
MQM	40.28%	34.97%	24.75%	6.50%

5.2.2 Twitter users with actual voters

In order to relate our finding with Top parties we have plotted party percentage with overall totals for these parties for twitter and actual votes Fig. 5.3. Here we have found some differences but they can be defended by valid arguments. In actual PMLN is on top for Top 5 Parties with PTI and PPP(P) on second and third place in terms of percentages. The twitter affiliation

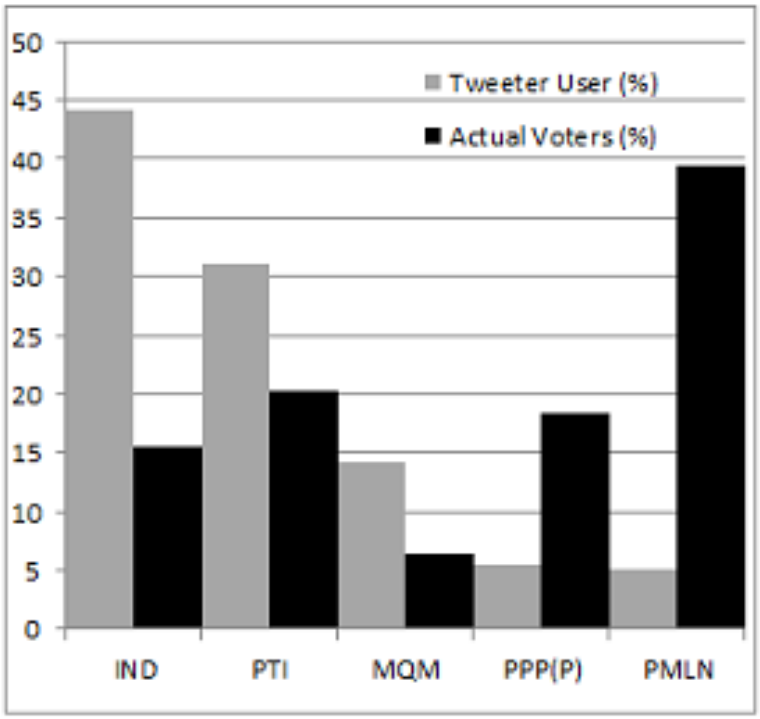

Figure 5.3: Twitter users with actual voters

showed PTI and MQM lead by around 44 independent twitter users. This major chunk of twitter user's may have created upside down situations for certain parties.

5.2.3 Twitter users versus Total votes polled

From Fig. 5.4, we can see presence of twitter user for Federal Capital and all four provinces. The plotted percentages showed that twitter was extensively used throughout Pakistan. If we relate total votes polled percentage with total twitter user percentage we can conclude that for this application was extensively used for Islamabad and Sindh (Karachi).

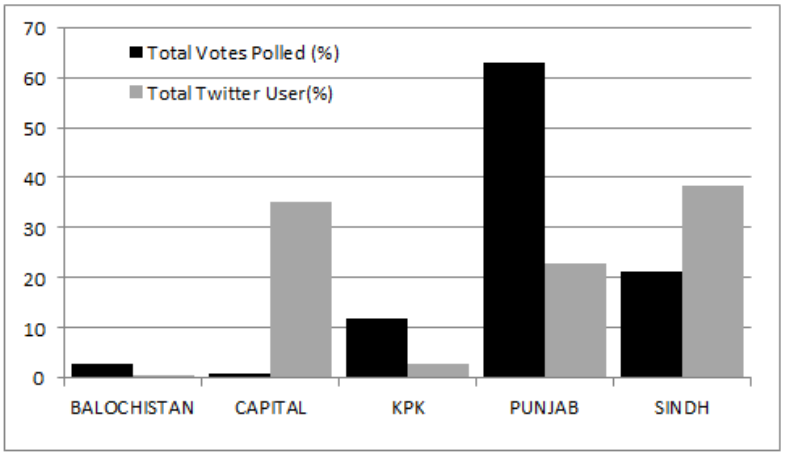

Figure 5.4: Twitter users versus Total votes polled.

Chapter 6

Conclusion and Future Work

In this chapter, the conclusion is drawn by presenting summary of the research findings along with future directions.

6.1 Conclusion

Sentiment detection / opinion mining from text obtained from social media had been appealing topic for Natural Language Processing. We tried to find out the predictive power of Twitter. By using different classifiers we achieved average 70% accuracy for positive and negative sentiment. We also achieved average 50% accuracy for manual labeled data for three classes positive, negative and neutral. With the help of these classifiers we classified political tweets related to Pakistan, which were used along other important entities for analysis. This work of forecasting elections for developing countries especially for 'Pakistan' is of unique attempt. With sentiment detection and analysis part we unveiled positive influence of Twitter users for Pakistan. We have found twitter users contributing for all political parties. We also deduced that there are certain political parties and leaders who have low electability but high popularity from Elections results in actual and our predictions. We have seen independent twitter users acted as drivers for change. We also ignored on ground realities of not considering major part of population living in rural areas with tribal loyalties. These people with low literacy rate plays vital role in setting political party's fortune as it could be seen in the case of PMLN and PPP(P) traditional rivals. We have seen, with the inclusion of PTI the left right alignment of these parties have been controlled to some extent. We also detected sentiment shift, as a spike in Fig. 4.1 could be seen. This spike showed rise in tweets rate especially for PTI with reason already discussed. Due to this fall PTI gain positive sentiment shift from its

supporter influencing final poll day results. In addition to above, the political parties may use twitter as a parameter for refining their campaign, and redefining their goals. We also found that no mechanism was adopted by ECP for stopping social media campaign over internet, twitter in our case, as it continued till and by voting date.

6.2 Future Work

As a future work, improvement of accuracy is primary concern by applying new techniques for feature selection and reductions. More efforts to be put in finding out some mechanism for inclusion of twitter users from rural areas over social media political campaign. More work has to be done in finding out users and their contents creating propaganda for election campaigns with identification of their political affiliations. Also using twitter data-set, we would like to predict desires of particular political party twitter followers regarding coalition with other parties before elections. How people think their future government should be like? What should be its mandate? Let's find out ways, if we could answer above mentioned questions and ideas.

Bibliography

[1] M. Skoric, N. Poor, P. Achananuparp, E. Lim, J. Jiang (2012), Tweets and Votes: A Study of the 2011 Singapore General Election, 45th Hawaii International Conference on System Sciences, Pages 2583-2591.

[2] A. Bermingham and A. F. Smeaton (2011), On Using Twitter to Monitor Political Sentiment and Predict Election Results, Sentiment Analysis where AI meets Psychology (SAAIP), Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), Chiang Mai, Thailand.

[3] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe (2010), Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, International AAAI Conference on Weblogs and Social Media.

[4] F. Nooralahzadeh, V. Arunachalam, C. Chiru (2013), 2012-Presidential Elections on Twitter-An Analysis of How the US and French Election were Reflected in Tweets, 19th International Conference on Control Systems and Computer Science.

[5] S. Stieglitz, L. D. Xuan, (2012), Political Communication and Influence through Microblogging-An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior, 45th Hawaii International Conference on System Sciences.

[6] X. Zhou, X. Tao, J. Yong, (2013), Sentiment Analysis on Tweets for Social Events, Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design.

[7] N. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, K. P. Gummadi, (2012), Inferring Who-is-Who in the Twitter Social Network, Workshop on Online Social Networks (WOSN)-ACM SIGCOMM.

[8] S. Alsaleem, (2011), Automated Arabic Text Categorization Using SVM and NB, International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011

[9] A. Choudhary, W. Hendrix, K. Lee, D. Palestia, W. K. Liao (2012), Social media evolution of the Egyptian revolution, Communications of the ACM, Vol. 55, No. 5, DOI:10.1145/2160718.2160736

[10] H. Bhavsar and A. Ganatra (2012), A Comparative Study of Training Algorithms for Supervised Machine Learning, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4.

[11] A. Boutet, H. Kim, E. Yoneki, (2012), What's in Twitter: I Know What Parties are Popular and Who You are Supporting Now!, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[12] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, F. Menczer, (2011), Predicting the Political Alignment of Twitter Users, IEEE, International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

[13] J. Chung and E. Mustafaraj (2011), Can Collective Sentiment Expressed on Twitter Predict Political Elections? Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence.

[14] B. O.'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith (2010), From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, Fourth International AAAI Conference on Weblogs and Social Media.

[15] D. G. Avello , P. T. Metaxas, E. Mustafaraj (2011), Limits of Electoral Predictions Using Social Media Data, Fifth International AAAI Conference on Weblogs and Social Media.

[16] P. Cogan, M. Andrews, M. Bradonjic (2012), Reconstruction and Analysis of Twitter Conversation Graphs, HotSocial '12 Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, Pages 25-31.

[17] A. Kumar and T. M. Sebastian (2012), Sentiment Analysis on Twitter, IJCSI International Journal of Computer Science.

[18] Twitter, `http://www.twitter.com`

[19] Twitter API, Wiki, `http://apiwiki.twitter.com/API-Overview`

[20] `https://pypi.python.org/pypi/tweetstream`

[21] M. Kaschesky, P. Sobkowicz, G. Bouchard (2011), Opinion Mining in Social Media: Modeling, Simulating, and Visualizing Political Opinion Formation in the Web, 12th Annual International Conference on Digital Government Research.

[22] P. W. Liang, B. R. Dai (2013), Opinion Mining on Social Media Data, 14th International Conference on Mobile Data Management.

[23] S. Asur, B. A. Huberman (2010), Predicting the Future With Social Media, International Conference on Web Intelligence and Intelligent Agent Technology.

[24] M. K. Dalal, M. A. Zaveri (2011), Automatic Text Classification: A Technical Review, International Journal of Computer Applications (0975 8887), Volume 28 No.2, August 2011

[25] A. L, M. P. Simmons, E. Adar, L. A. Adamic (2011), The Party is Over Here: Structure and Content in the 2010 Election, Fifth International AAAI Conference on Weblogs and Social Media.

[26] J. Han, M. Kamber and J. Pei (2012), Data Mining:Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers, ©Elsevier Inc.

[27] Pildat, `http://www.pildat.org`

[28] K. Andreas, H. Michael, (2010), Users of the world, uniteThe challenges and opportunities of social media, Business Horizons, Vol. 53, Issue 1 (page 61).

[29] Election Commission of Pakistan, `http://www.ecp.gov.pk`

[30] N. K. Chebib and R. M. Sohail (2011), The reasons social media contributed to the 2011 Egyptian revolution, International journal of business research and management (IJBRM).

[31] C D. Manning, P Raghavan and H Schtze (2008), Introduction to Information Retrieval, Cambridge University Press.