

CHAPTER 1

INTRODUCTION

1.1 Overview

As twitter boasts of the sentiments of the people in real time about various issues, and elections being a game of sentiment, we are trying to detect sentiment for a party or a candidate in general elections, 2014.

1.2 Motivation

Nowadays, social networks such as Twitter are the latest trend in the globalized word. With more than 40 million users using the 140-characters microblogging platform, Twitter became a process with its own dynamic. It is used in different scenarios by a broad set of different users. Each of them has their own behavioural pattern, their own style of writing. Mining these platforms may extract valuable information. Currently, there is heavy research going on in this area, with promising results. Also, Twitter was used to monitor the U.S. presidential debate in 2008 .Tweeters tended to favour Obama over McCain, and Obama really won the election afterwards. This shows that Twitter can also be used to predict political election results.

With this in mind, we want to apply state-of-the-art classification tools to extract information from Twitter. We want to answer the question if it is possible to automatically classify Tweets (and thus the users themselves). This may help to separate professionally created news messages from user-generated messages. News messages tend to have an informational character, while user-generated messages usually have a communicational purpose.

1.3 Sentiment Analysis

-What is Sentiment Analysis?

Sentiment analysis is an application of text analytics techniques for the identification of subjective opinions in text data. It normally involves the classification of text into categories

such as "positive", "negative" and in some cases "neutral". Over the last five years, there is an increasing demand for sentiment analysis tools by companies willing to monitor customer's or employee's opinions of the company and on its products and services. To fulfill the increasing demands for such kinds of tools, more and more researchers and companies are releasing products to perform sentiment analysis, many of them claiming to be able to perform sentiment analysis of any type of document in every domain. Unfortunately, experience has shown us that, an "out-of-the-box" sentiment analysis tools working across domains does not yet exist. The main reason sentiment analysis is so difficult is that words often take different meanings and are associated with distinct emotions depending on the domain in which they are being used. There are even situations where different forms of a single word will be associated with different sentiments. For example, in customer feedback that the word "improved" was associated with positive comments, but "improve" was more often used in negative ones. All sentiment analysis tools rely, at varying degrees, on lists of words and phrases with positive and negative connotations or are empirically related to positive or negative comments.

1.4 Why Sentiment is hard to decipher?

Online comments don't fall neatly into "positive" and "negative" buckets. There's a range of consumer sentiment that challenges even the most sophisticated natural language processing technologies. In a Sentiment Analysis Symposium, Catherine van Zuylen, VP of products at Attensity, a social analytics software vendor, provided this list of difficult comment-analysis problems:

- **False negatives:** The words "crying" and "crap" suggest negativity, but then there is "I was crying with joy" or "Holy crap! This is great." Here's where simplistic tools might be fooled.
- **Relative sentiment:** "I bought a Honda Accord.", which is great for Honda but bad for Toyota.
- **Compound sentiment:** Doing work for movie studies. Such as "I loved the trailer but hated the movie." Big mobile phone companies encounter mixed messages such as "I love the phone but hate the network."
- **Conditional sentiment:** "If someone doesn't call me back, I'm never doing business with them again." Or "I was really pissed, but then they gave me a refund."

- **Scoring sentiment:** Vendors are expected to measure relative sentiment, but how positive is "I like it" versus "I really like it" versus "I love it"?
- **Sentiment modifiers:** "I bought an iPhone today :-)" or "Gotta love the cable company ;-<". Emoticons are straightforward, but what words are they connected to?
- **International sentiment:** Japanese have unique emoticons, like (;_;) for crying. Italians tend to be far more effusive and grandiose, whereas Brits are generally drier and less effusive, making those relative scoring challenges mentioned earlier all the more complicated.

1.5 Application in Political Domain

Twitter sentiment analysis has revealed great amount of information in the previous elections and other big event occurring in the world. Using the analysis of the twitter messages, i.e. whether a particular topic or a product is trending in a positive, negative or a neutral way, market researchers have a valuable tool to monitor how a product is accepted, or how a phenomenon is taken up by the general public. This classification can help market researchers know the sentiment towards a product or an individual personality which can thus help in taking decisions based on large number of user views. Examples of this include during the 2008 US elections presidential debates, Obama trended very high as compared to the other contender and in the end Obama won. So, like this many such predictions could also be done. This report focuses on the Indian general elections occurring during the Apr-May, 2014.

CHAPTER 2

BACKGROUND AND LITERATURE SURVEY

2.1 About Twitter

The Internet has become a global forum where people express their opinions. With the recent explosive growth of the social media content on the Web, people post reviews of products on merchant sites and express their views about almost anything in their personal blogs, pages at social network sites like Facebook, Twitter, and Blogger. People are also eager to know what individual journalists, and columnists are thinking or feeling about subjects such as politicians, political parties and social issues. With the rapid growth in news groups it is possible to analyze opinions and feelings in news domain, too. For these reasons, sentiment analysis aiming to determine the attitude (sense, emotion, opinion etc.) of a speaker or a writer with respect to a specified topic has become a major area of interest in the field of NLP. Various statistical and linguistic methods have been developed for the Sentiment Analysis of English texts for different domains.

2.1.1 What is twitter?

Twitter is a popular social network site which only asks one question: 'What are you doing?'. The answer is limited to 140 characters. Figure 2.1 shows a screenshot of the current Twitter User Interface. Status updates can be sent via a web browser, SMS, e-mail or third party applications and are displayed on the users' profile.

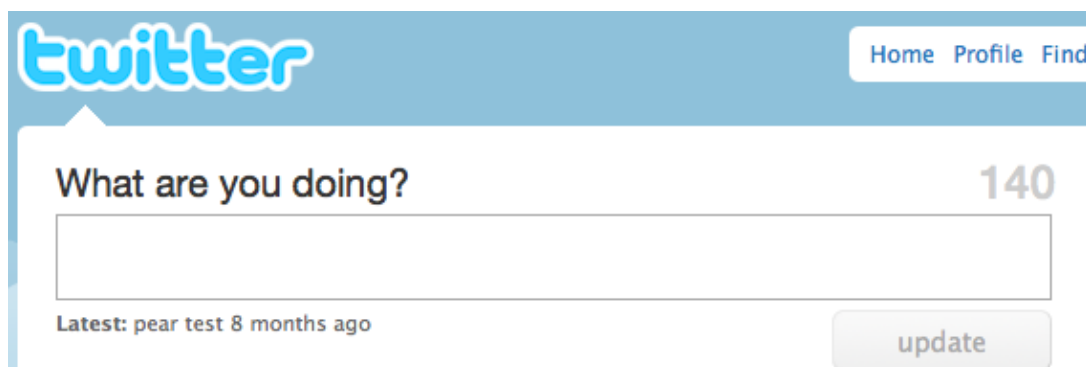


Fig 2.1 : twitter.com UI to compose a new tweet

2.1.2 Followers

Twitter implemented a concept of so-called followers. If a certain user updates his/her status, all followers are informed of the new status. This is achieved by adding the new entry to their personal Twitter overview page

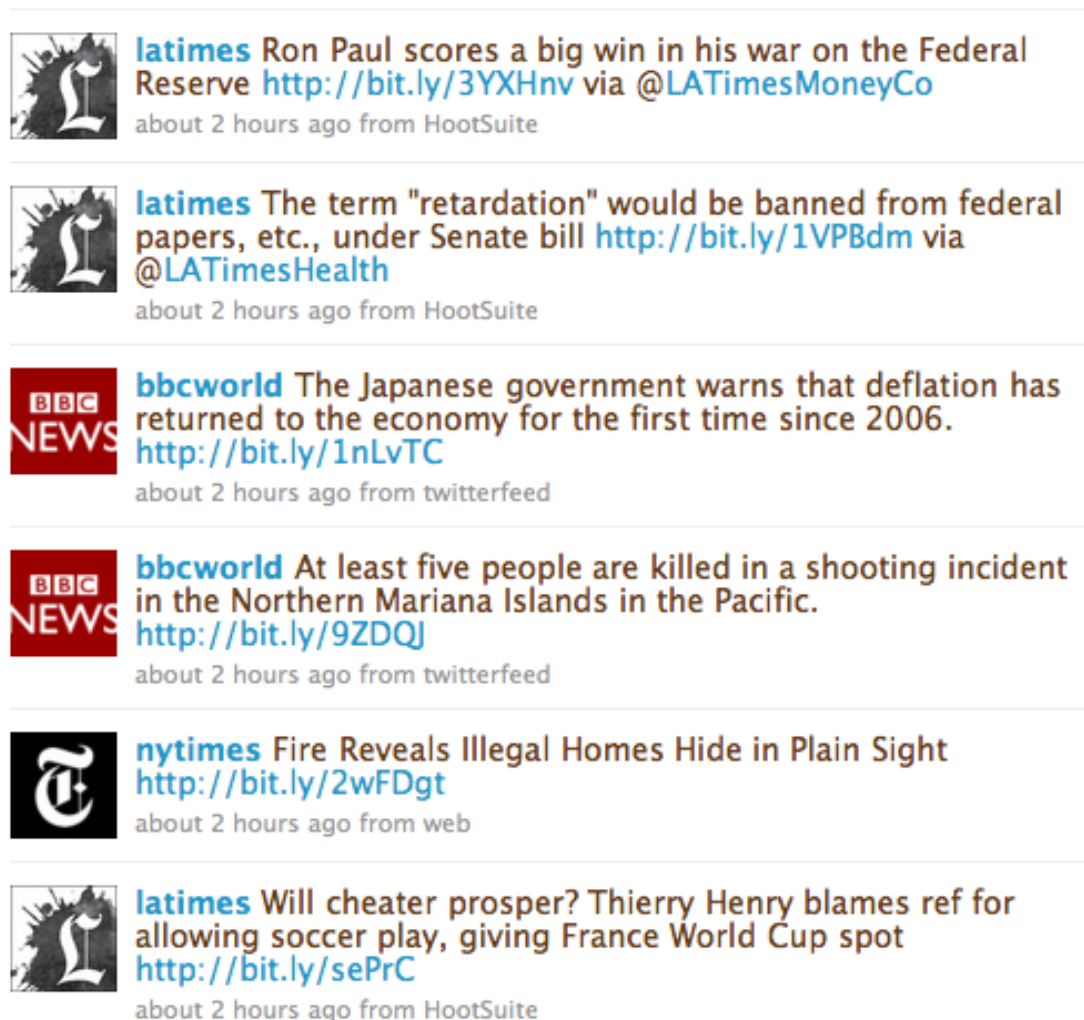


Fig 2.2 Screenshot of the personal Twitter overview page

2.1.3 Business Model

Like several other popular social network sites, Twitter struggles to find a valid business model which actually generates revenue. Twitter itself confirms this on their web page:

“Twitter has many appealing opportunities for generating revenue but we are holding off an implementation for now because we don't want to distract

ourselves from the more important work at hand which is to create a compelling service and great user experience for millions of people around the world. While our business model is in a research phase, we spend more money than we make.¹ To finance the service Twitter relies heavily on investors and has thus generated a total funding of 155 Million dollars. According to the Financial Times, the investors currently value the site with 1 Billion dollars.”

2.1.4 Twitter Slang

Due to the 140 characters limit, users have developed strategies to put as much information as possible in the messages. This includes the usage of the hash character (#) to tag a message with certain topics. For example, a message may look like this: ‘I’m currently testing a new Twitter feature. #test #twitter #graz’. The message is now tagged with ‘test, twitter, graz’, and other people are able to search for that tags using the Twitter site. Also, the usage of URL shortening services became very popular. Those services allow shortening a certain URL so that it can be posted on Twitter. For example, <http://www.know-center.tugraz.at/forschung/> becomes <http://tinyurl.com/yh8lqhz> using the shortening-service ‘TinyURL.com’. TinyURL.com was the first well-established shorting services and thus was the default service in Twitter. But in early 2009 Twitter silently switched to bit.ly for unknown reasons.

The New York Times reports that

“Bit.ly, which recently raised \$2 million in venture financing, tracks real-time statistics on how many times links are clicked and where users are coming from information that could be valuable to companies and brands looking to measure the impact of an e-mail message, ink, tweet or mention online.”

2.1.5 Geographical distribution

Java et al. carried out a detailed analysis of Twitter in the year 2007, being one of the first scientific papers to deal with this topic. They performed a detailed geographical analysis. The results show that Twitter is mostly used in the United States (especially East Coast), in

Europe and Asia (mainly Japan). Twitter is adopted the most in the cities Tokyo, New York and San Francisco. Figure 3.5 visualizes the geographical distribution. Also, Java et al. found out that the social network ties are closer in Europe and Asia than in North America.

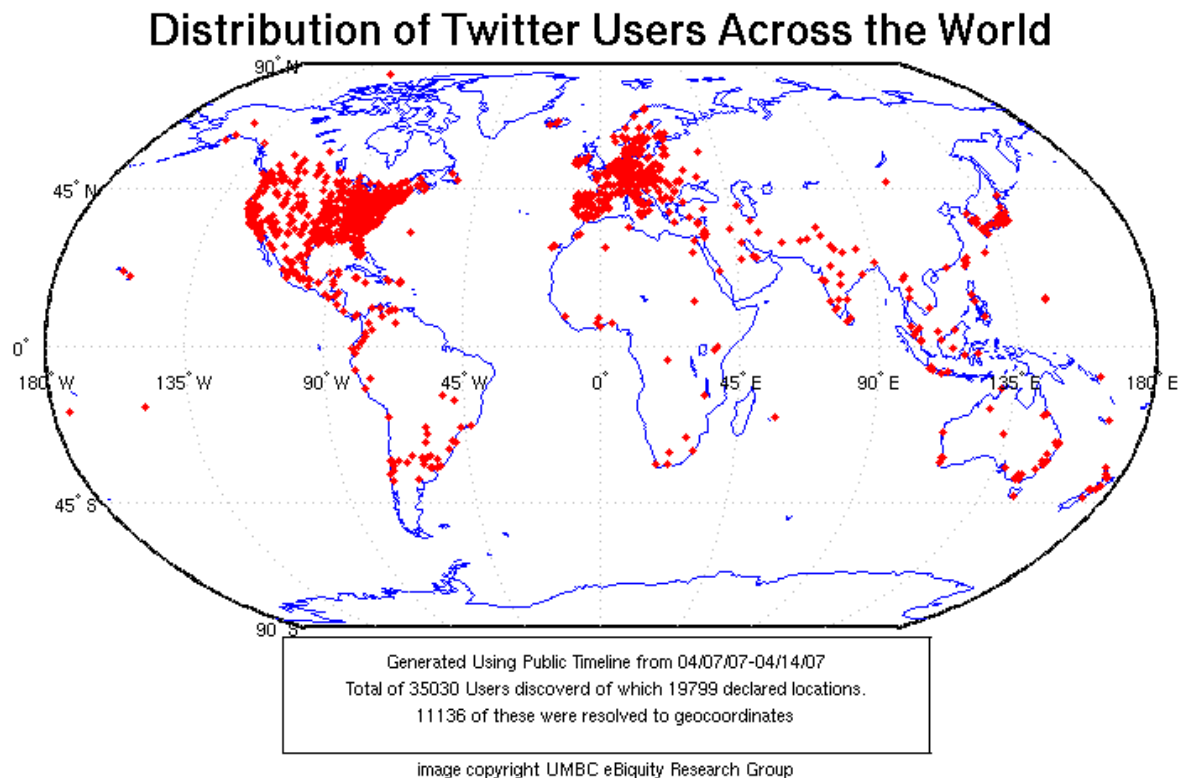


Fig 2.3: Distribution of Twitter users across the World

2.2 Political usage

During the 2009 election in Iran, Twitter played an important role. While the newspapers and blogs in Iran were heavily censored by the government, Twitter users published news from the street in real-time. They either used the hashtags #iran or #iranelection. News cooperation from all over the world displayed the latest Twitter messages. After a while, the government tried to suppress those messages. Users reacted and asked all Twitter users to change their location to 'Teheran, Iran', making it impossible for the Iranian Ministry of Intelligence to locate those people.

The U.S. State Department knew about the importance of this communication tool and asked Twitter to delay a scheduled upgrade which took place 3 days after the election on

June 15th, 2009. Twitter agreed and carried out the upgrade on 2pm U.S. time, which is 1.30am Teheran time.

The leader of the opposition, MirHossein Mousavi, even twittered his arrest: Dear Iranian People, Mousavi has not left you alone, he has been put under house arrest by Ministry of Intelligence #IranElection 3 Also, Twitter was used to raise funds for the victims of the Haitian earthquake. Shortly after the earthquake, the American Red Cross sent following Tweet: 'You can text 'HAITI' to 90999 to donate \$10 to Red Cross relief efforts in #haiti.'

Soon, the famous Haitian singer Wycli Jean⁵ started to support this campaign, and several other celebrities followed him. This shows that Twitter is playing an important role in the society these days and can be used for more than just ordinary status updates of what you're currently doing.

2.3 Data Mining

Data mining refers to extracting or “mining” knowledge from large amounts of data. It can also be named by “knowledge mining form data”. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/ pattern analysis, data archaeology, and data dredging.

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases, or KDD. Alternatively, data mining is also treated simply as an essential step in the process of knowledge discovery in databases.

The fast-growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. In such situation we become data rich but information poor. Consequently. Important decisions are often made based not on the information-rich data stored in databases but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. In addition, current expert system technologies rely on users or domain experts to manually input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time consuming and costly. In such situation data mining tools can perform data analysis and may uncover important data patterns, contributing greatly to business strategies,

knowledge bases, and scientific or medical research.

2.4 Machine Learning

Machine Learning is defined as "the ability of a machine to improve its performance based on previous results". In other words it is a system capable of learning from experience and analytical observation, which results in continuous self improvement there by offering increased efficiency and effectiveness. In general there are four different types of machine learning techniques. They are:

1. Supervised learning.
2. Unsupervised learning.
3. Semi-supervised learning and
4. Reinforcement learning.

And this project deals with text categorization which is a supervised learning technique.

2.4.1 Supervised Learning

Supervised learning is a machine learning technique that learns from training data set. A training data set consists of input objects, and categories to which they belong. Assigning categories to input objects is carried out manually by an expert. Given an unknown object, supervised learning technique must be able to predict an appropriate category based on prior training.

2.5 Summary of relevant papers

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009). Microblog data like Twitter, on which users post real time reactions to and opinions about “everything”, poses newer and different challenges.

Paper 1:

Title of paper-Analysis and Classification of Twitter messages

Authors-Christopher Horn, Dipl.-Ing,Dr.techn. Michael Granitzer

Year of Publication-April 2010

Publishing Details- Resolution of the Curriculum Commission for bachelor's, master's and diploma programs

Summary-

In this thesis, a way for analyzing Twitter messages (and as a result, the users) was proposed. It was necessary that the representatives do not overlap each other and they constitute an optimal representative for their category. After defining and specifying the categories, author shave manually chosen 120 Twitter users and assigned them to a category i.e. 'User' Tweets, News, Company Advertisements. In regard to sentiment detection, a scoring approach using a list of positive and negative words was used. The scientific output of this paper is a dataset containing 120 users and approximately 4800 Tweets, all categorized. Also, a platform was implemented which can be used to classify arbitrary Twitter users. And last but not least, our empirical evaluation showed that Support Vector Machines can be used very well for classifying Twitter messages. The most challenging task would be to use the proposed scoring algorithm to automatically bootstrap the training examples for the SVM - thus switching to a semi-supervised learning setting - and examine if the created system outperforms current sentiment detectors. Also, it would be interesting to implement different classifiers (like kNN, or Random Forests) and compare the results.

Paper 2:

Title of paper-Sentiment Analysis of Twitter Data

Authors-ApoorvAgarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau

Year of Publication-Summer 2013

Publishing Details-Proceeding, LSM '11 Proceedings of the Workshop on Languages in Social Media, Pages 30-38, Association for Computational Linguistics Stroudsburg, PA, USA

Summary-

In this paper, authors had tried to compare the result of sentimental analysis of tweets using different algorithms like POS-specific prior polarity features, Kernel Tree, Naive Bayes, MaxEnt and SVM. They have also written about classification of tweets in binary or 3-tier category. To build a training set, they collected data and annotated it manually as positive, negative or neutral. Afterwards, they divided the annotated data into 2 buckets: training and testing data. The testing data is then preprocessed by discarding the junk (slag words). The emoticons, hyperlinks, hash tags, targets, acronyms etc. are cleaned. For this, they have referred to dictionary or list of particulars. Using this training data, testing data was analyzed and checked if the annotation result and computed result is same or not. They concluded that sentiment analysis for Twitter data was not that different from sentiment analysis for other genres.

Paper 3:

Title of paper-Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing

Authors-Elena Filatova

Year of Publication-23-25 May, 2012

Publishing Details-Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12;) ISBN = {978-2-9517408-7-7}

Summary-

In this paper author describe a corpus generation experiment whose goal is to obtain Amazon product review pairs that can be used to analyze the notion of sarcasm in text and for training sarcasm detection system. Our corpus can be used for understanding sarcasm on two levels: document and text utterance level. The presence of irony in a product review does not affect the readers' understanding of the product quality: given the text of the review (irrespective, whether it is a sarcastic or a regular review), people are good in understanding the attitude of the review author to the product under analysis and can reliably guess how many stars the review author assigned to the product.

CHAPTER 3

DESIGN

3.1 Data Gathering

The process begins with collecting tweets from the public stream of twitter.com . After creating an account and generating an OAuth key, we can stream tweets in our application, using various search queries. For this project, we are trying to estimate support for the most famous leaders on twitter for various political parties. Using this, we are trying to estimate how many users are supporting which political party, and how actively the supporters are tweeting for their favourite leaders. The various search terms used for tweet collection include “modi”, “rahul gandhi”, “arvind kejriwal”, “bjp”, “congress”, “aap”. These keywords have been used by monitoring twitter over various periods, and for having best results, we have used equal keywords for each party. Although in indian election sphere, aap is a newbie, and would have little impact in rural areas, therefore we are only estimating the support of the users in urban population, where twitter is more active as compared to semi-urban or rural areas.

The structure of the tables used for storing data are:

```
>desc tweetcollection;
```

Field	Type	Null	Key	Default	Extra
username	varchar(30)	YES		NULL	
profileLocation	varchar(50)	YES		NULL	
tweetId	bigint(20)	NO	PRI	0	
status	varchar(200)	YES		NULL	

Table 3.1: Structure of Tweet storage

3.2 Data Cleaning

The most important factor for the quality of the result would be how better the data which we have collected is. So various processes that are applied for cleaning the data are as follows:

3.2.1 Removing website links

Tweets have website links embedded in them, and in order to create a good quality data set, we will remove these links.

For e.g. consider the following tweet

“ BJP's PM candidate Narendra Modi addresses BJP supporters at a rally in Lakhimpur, Uttar Pradesh #India2014 <http://t.co/WDm5bQbeK> “.

“In a 2nd attack on him in 4 days, Arvind Kejriwal was slapped by a man during a road show in Delhi's Sultanpuri area <http://t.co/4Hq3Lm7F7A> “.

“FIR registered against Beni Prasad Verma for a remark he made against Narendra Modi & Rajnath Singh <http://t.co/QkFfLcj5iv> “

The http removed tweets are:

“ BJP's PM candidate Narendra Modi addresses BJP supporters at a rally in Lakhimpur, Uttar Pradesh #India2014”

“In a 2nd attack on him in 4 days,Arvind Kejriwal was slapped by a man during a road show in Delhi's Sultanpuri area”

“FIR registered against Beni Prasad Verma for a remark he made against Narendra Modi & Rajnath Singh”

3.2.2 Removing # and @

For removing the # and @, we have the following examples:

“@Real_Ashok: PM and Congress cannot take the Nation for Granted: Narendra Modi at 3D Rally <http://t.co/OIxcCPOz1O> “

“@narendramodi_in: The world will see...on 16th May all those who ruined India will be defeated: Narendra Modi <http://t.co/FsxK4LRG1I> “

And after removing are:

“PM and Congress cannot take the Nation for Granted: Narendra Modi at 3D Rally”.

“The world will see...on 16th May all those who ruined India will be defeated: Narendra Modi”.

3.2.3 Removing stopwords

Stopword create a noise in the dataset as they are present in all the category of tweets, whether positive or negative. Following examples show the removal.

“Congress Vice President Rahul Gandhi addresses a rally in Raichur, takes a jibe at Narendra Modi's Gujarat model”

“Taking on Narendra Modi, Rahul Gandhi says BJP's balloon will soon burst”

And after removing are:

“Congress Vice President Rahul Gandhi addresses rally Raichur, takes jibe Narendra Modi's Gujarat model”

“Taking Narendra Modi, Rahul Gandhi says BJP's balloon soon burst”

3.2.4 Removing special characters

“Ramdev apologises for `honeymoon` remark on Rahul Gandhi, BJP gives mixed reactions: A day after his controversy.”

“RT @upma23: Good sign that 'evil' empire of Sonia Gandhi and her cronies are crumbling...”

And after removing are:

“Ramdev apologises honeymoon remark Rahul Gandhi, BJP gives mixed reactions day after his controversy”

“Good sign that evil empire of Sonia Gandhi her cronies crumbling”

3.3 Data Classification

Classifying tweets into positive and negative is done using LingPipe. LingPipe is tool kit for processing text using computational linguistics.

The system makes use of LingPipe toolkit to obtain the classifier by initially providing hand classified tweets as positive or negative for the 3 political parties, namely BJP, Cong and AAP. We provided 10 manually classified tweets in their folder, and created the classifier using this set. This process was repeated for all the 3 parties. Now we selected 1000 random tweets for each party using search query in MySQL to create table for 1000 tweets for each party using the following query keyword.

Now, after creating the classifier using first 1000, divided into positive and negatives for each party, the classifier was now again created using this bigger training data set. After this, we used the complete data set of ~2.5 lakh tweets (0.25 Million) and used this as our training data and classified them as positive or negative for each party.

Party	Keywords
BJP	'%modi%', '%bjp%'
Cong	'%cong%', '%gandhi%'
AAP	'%aap%', '%kejriwal%'

Table 3.2: Keywords used for Tweets collection

3.4 Technology Used

This project is a web application that is developed using the following softwares:

1. NetBeans IDE 7.1.1
2. JDK 7
3. MySQL
4. LingPipe
5. Twitter API
6. Twitter4j

3.4.1 NetBeans IDE 7.1.1

As per the information provided by Wikipedia, NetBeans is an integrated development environment (IDE) for developing primarily with Java, but also with other languages, in particular PHP, C/C++, and HTML5. It is also an application platform framework for Java desktop applications and others.

The NetBeans IDE is written in Java and can run on Windows, OS X, Linux, Solaris and other platforms supporting a compatible JVM.

The NetBeans Platform allows applications to be developed from a set of modular software components called modules. Applications based on the NetBeans Platform (including the NetBeans IDE itself) can be extended by third party developers.

3.4.2 JDK7

The Java Development Kit (JDK) is an implementation of either one of the Java SE, Java EE or Java ME platforms released by Oracle Corporation in the form of a binary product aimed at Java developers on Solaris, Linux, Mac OS X or Windows. JDK 7 is a version of JDK.

3.4.3 MySQL

MySQL is (as of March 2014) the world's second most widely used open-source relational database management system (RDBMS). It is named after co-founder Michael Widenius's daughter. The SQL phrase stands for Structured Query Language.^[6]

The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.

MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack. LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Free-software-open source projects that require a full-featured database management system often use MySQL.

3.4.4 LingPipe

LingPipe is tool kit for processing text using computational linguistics. LingPipe is used to do tasks like:

- Find the names of people, organizations or locations in news
- Automatically classify Twitter search results into categories
- Suggest correct spellings of queries

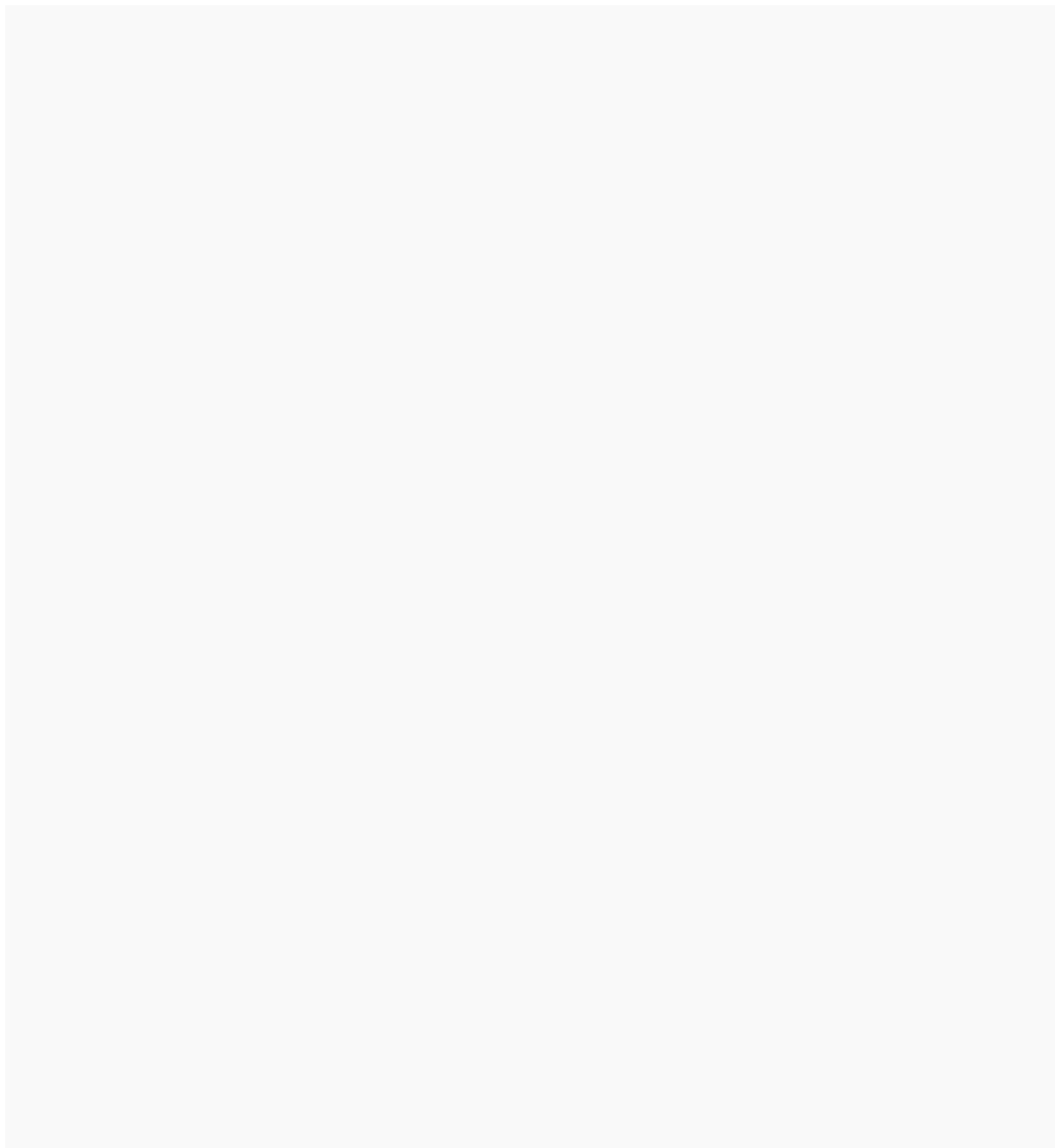
3.4.5 Twitter API

Twitter API has been created by twitter.com and is their official API. They are used to provide access to various twitter activities like posting a new tweet, getting new tweets for a

user or getting tweets from its public stream for various statistical and analysis purposes. It provides authentication keys used for creating token for different privileges when a developer creates an account.

3.4.6 twitter4j

Twitter4J is an unofficial Java library for the Twitter API. With Twitter4J, you can easily integrate your Java application with the Twitter service. Twitter4J is an unofficial library.



CHAPTER 4

IMPLEMENTATION

4.1 Overall Description and Design

We are trying to detect sentiment of twitter messages using the LingPipe library, to classify them as either positive or negative.

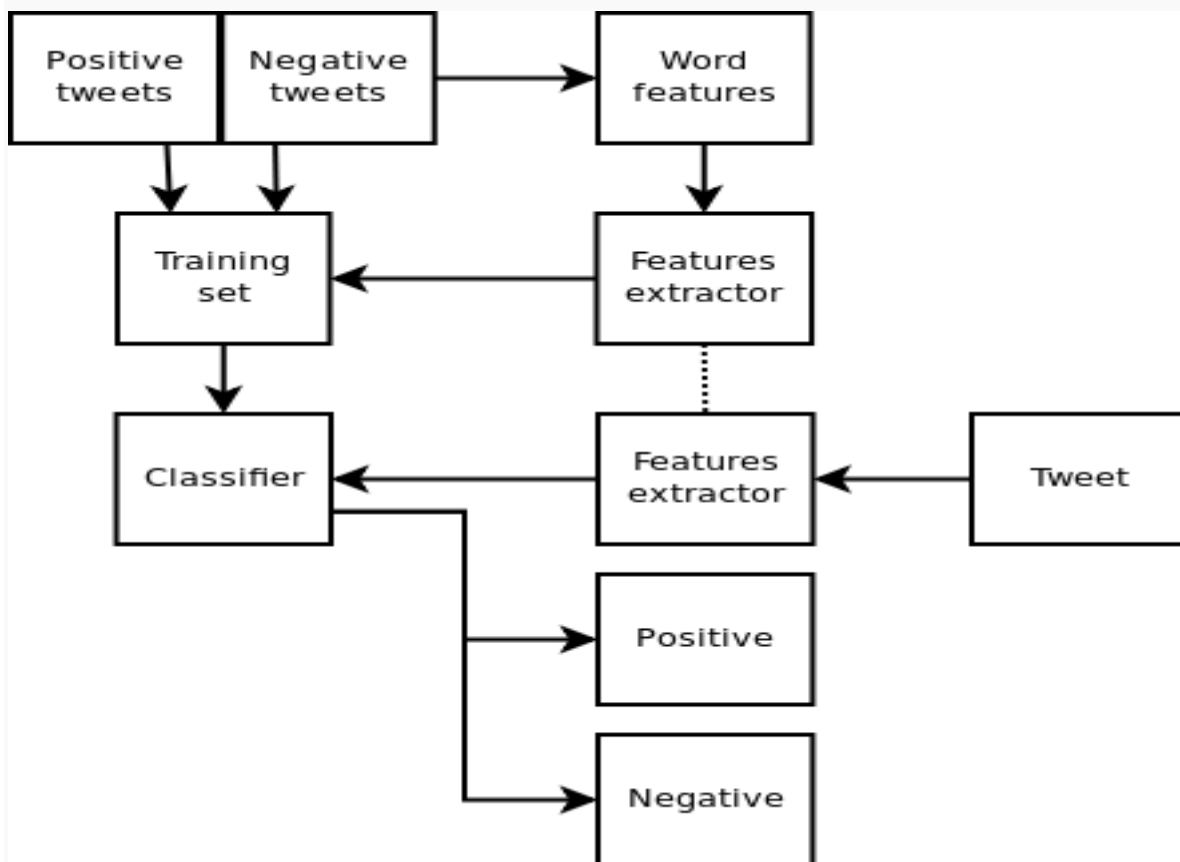


Fig 4.1: Way to Classify Tweets

4.2 Requirement Specification

4.2.1 Getting the tweets using public stream of twitters

The first requirement of our system is to stream tweets from twitter.com, using the twitter API and using twitter4j library. The code for this is given below

```

import twitter4j.FilterQuery;
import twitter4j.StallWarning;
import twitter4j.Status;
import twitter4j.StatusDeletionNotice;
import twitter4j.StatusListener;
import twitter4j.TwitterStream;
import twitter4j.TwitterStreamFactory;
import twitter4j.User;
import twitter4j.conf.ConfigurationBuilder;

public class SimpleStream {
    Connection conn;

    public static void main(String[] args) {
        final SimpleStream obj=new SimpleStream();
        ConfigurationBuilder cb = new ConfigurationBuilder();
        cb.setDebugEnabled(true);
        cb.setOAuthConsumerKey("zF6yTLcFfz0Digt18NYow");
        cb.setOAuthConsumerSecret("NA5ASnUcn5hDOOFTKzQCoCbqpauAtsEOxnLkz6kNpVk");
        cb.setOAuthAccessToken("2328489986-
FrKbIaQPjJ27Fu9cj4W6pgCzXGRQnuV34yS4m1");
        cb.setOAuthAccessTokenSecret("JWQ5LzU3lgN2gYdZ8V508uNulDsy8K4UNdquKv2uP1quQ
");

        final SentimentClassifier sentClassifier=new SentimentClassifier();
        StatusListener listener = new StatusListener() {

            @Override
            public void onException(Exception arg0) {
            }

            @Override
            public void onDeletionNotice(StatusDeletionNotice arg0) {
            }

            @Override
            public void onScrubGeo(long arg0, long arg1) {
            }
        }
    }
}

```

```

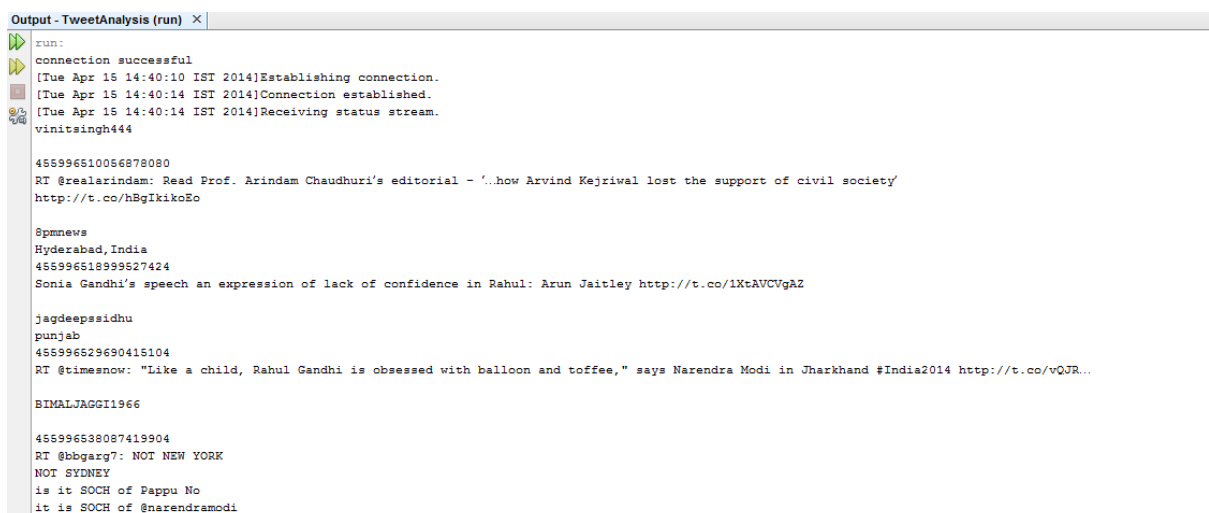
@Override
public void onStatus(Status status) {
    User user = status.getUser();
    // gets Username
    String username = status.getUser().getScreenName();
    System.out.println(username);
    String profileLocation = user.getLocation();
    System.out.println(profileLocation);
    long tweetId = status.getId();
    System.out.println(tweetId);
    String content = status.getText();
    System.out.println(content + "\n");

    FilterQuery fq = new FilterQuery();

    String keywords[] = {"#bjp", "#aap", "#cong", "namo", "rahul
gandhi", "kejriwal"};
    fq.track(keywords);
    twitterStream.addListener(listener);
    twitterStream.filter(fq);

}
}

```



```

Output - TweetAnalysis (run) X
run:
connection successful
[Tue Apr 15 14:40:10 IST 2014]Establishing connection.
[Tue Apr 15 14:40:14 IST 2014]Connection established.
[Tue Apr 15 14:40:14 IST 2014]Receiving status stream.
vinitsingh444

455996510056878080
RT @realarindam: Read Prof. Arindam Chaudhuri's editorial - '...how Arvind Kejriwal lost the support of civil society'
http://t.co/hBgIkikoEo

8pmnews
Hyderabad,India
455996518999527424
Sonia Gandhi's speech an expression of lack of confidence in Rahul: Arun Jaitley http://t.co/1XtAVCVgAZ

jagdeepssidhu
punjab
455996529690415104
RT @timesnow: "Like a child, Rahul Gandhi is obsessed with balloon and toffee," says Narendra Modi in Jharkhand #India2014 http://t.co/vQJR...

BIMALJAGGI1966

455996538087419904
RT @bbqarg7: NOT NEW YORK
NOT SYDNEY
is it SOCH of Pappu No
it is SOCH of @narendramodi

```

Fig 4.2 Streaming of Tweets

```
mysql> select count(*) from tweetcollection2;
+-----+
| count(*) |
+-----+
|    258964    |
+-----+
1 row in set (0.06 sec)

mysql> _
```

Fig 4.3 Total Tweet Count

4.2.2 Cleaning of tweet implementation

Various functions applied for cleaning include

4.2.2.1 Removing http links

```
String removehttp(String str_received)
{
    String[] statusarray=str_received.split(" ");
    String str_tosend=new String();
    for(int i=0;i<statusarray.length;i++)
    {
        if(statusarray[i].startsWith("http://"))
        {
            System.out.println();
            System.out.println("removed http from "+(i+1)+"th word... :) ");
            str_tosend+=" ";
        }
        else
        {
            str_tosend+=statusarray[i];
            str_tosend+=" ";
        }
    }
}
```

```

    }
    return str_tosend;
}

```

4.2.2.2 Removing Stopword code

```

String removestopword(String str_received)
{
    String str_send=new String();
    String stopWords[] = { "a", "about", "above", "above",
    "across", "after", "afterwards", "again", "against", "all",
    "almost", "alone", "along", "already", "also", "although",
    "always", "am", "among", "amongst", "amoungst", "amount", "an",
    "and", "another", "any", "anyhow", "anyone", "anything", "anyway",
    "anywhere", "are", "around", "as", "at", "back", "be", "became",
    "because", "become", "becomes", "becoming", "been", "before",
    "beforehand", "behind", "being", "below", "beside", "besides",
    "between", "beyond", "both", "bottom", "but", "by", "call",
    "can", "cannot", "cant", "co", "com", "con", "could", "couldnt", "cry",
    "de", "describe", "detail", "do", "done", "down", "due", "during",
    "each", "eg", "eight", "either", "eleven", "else", "elsewhere",
    "empty", "enough", "etc", "even", "ever", "every", "everyone",
    "everything", "everywhere", "except", "few", "fifteen", "fify",
    "fill", "find", "fire", "first", "five", "for", "former",
    "formerly", "forty", "found", "four", "free", "from", "front", "full",
    "further", "get", "give", "go", "had", "has", "hasnt", "have",
    "he", "hence", "her", "here", "hereafter", "hereby", "herein",
    "hereupon", "hers", "herself", "him", "himself", "his", "how",
    "however", "hundred", "ie", "if", "in", "inc", "indeed",
    "interest", "into", "is", "it", "its", "itself", "keep", "last",
    "latter", "latterly", "least", "less", "ltd", "made", "many",
    "may", "me", "meanwhile", "might", "mill", "mine", "more",
    "moreover", "most", "mostly", "move", "much", "must", "my",
    "myself", "name", "namely", "neither", "net", "never", "nevertheless",
    "next", "nine", "nobody", "none", "noone",
    "nothing", "now", "nowhere", "of", "off", "often", "on", "once",
    "one", "only", "onto", "or", "org", "other", "others", "otherwise", "our",

```

"ours", "ourselves", "out", "over", "own", "part", "per",
 "perhaps", "please", "put", "rather", "re", "same", "see", "seem",
 "seemed", "seeming", "seems", "serious", "several", "she",
 "should", "show", "side", "since", "sincere", "six", "sixty", "so",
 "some", "somehow", "someone", "something", "sometime", "sometimes",
 "somewhere", "still", "such", "system", "take", "ten", "than",
 "that", "the", "their", "them", "themselves", "then", "thence",
 "there", "thereafter", "thereby", "therefore", "therein",
 "thereupon", "these", "they", "thick", "thin", "third", "this",
 "those", "though", "three", "through", "throughout", "thru",
 "thus", "to", "together", "too", "top", "toward", "towards",
 "twelve", "twenty", "two", "un", "under", "until", "up", "upon",
 "us", "very", "via", "was", "we", "well", "were", "what",
 "whatever", "when", "whence", "whenever", "where", "whereafter",
 "whereas", "whereby", "wherein", "whereupon", "wherever",
 "whether", "which", "while", "whither", "who", "whoever", "whole",
 "whom", "whose", "why", "will", "wikipedia", "with", "within", "without",
 "would", "yet", "you", "your", "yours", "yourself", "yourselves", "rt", "RT", "???",

"the", "????", "??", "???", "?????", "???????", "?????????", "???????????", "?????????????", "???????????????" };

```
String[] statusarray=str_received.split(" ");
boolean isastopword;
boolean isagoodword;

for(int i=0;i<statusarray.length;i++)
{
    isastopword=false;
    isagoodword=false;
    for(int j=0;j<stopWords.length;j++)
    {
        if((statusarray[i]).equalsIgnoreCase(stopWords[j]))
        {
            isastopword=true;
            isagoodword=false;
            break;
        }
    }
}
```



```

else{
    isagoodword=true;
}

}

if(isagoodword&&(!isastopword))
{
    str_send+=statusarray[i];
    str_send+=" ";
}
else{
    str_send+="";
}

}

return str_send; }

```

4.2.3 To define sentiments as Positive or Negative

This is the second part of our system. We need to develop an approach by which we can categorize the sentimental into positive or negative for different political parties. With this we will be able to easily make a judgment.

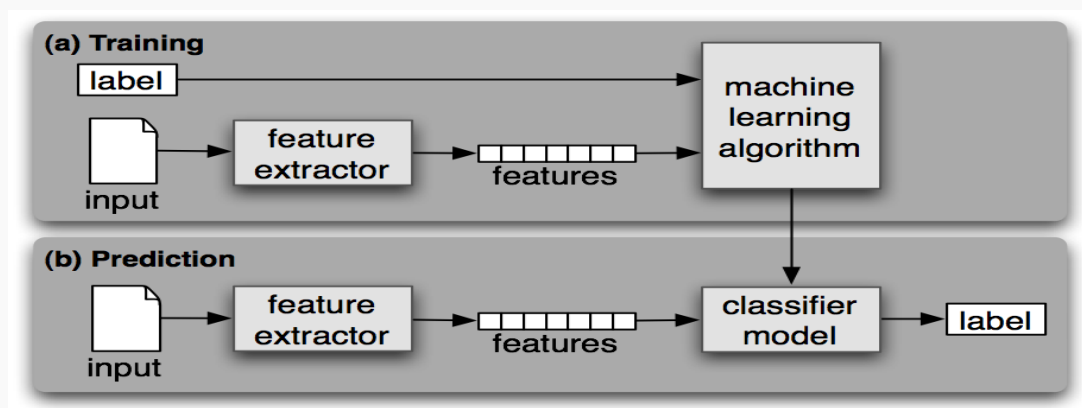


Fig 4.4 :Training and Testing Data

4.2.4 Feature Selection

Being one of the most important tasks in text classification, feature selection is the task of selecting the terms which are to be used in the training set. Selecting the terms has two advantages: A reduced term size speeds up the training process. Also, it reduces the noise by removing irrelevant features, thus increasing the classification accuracy. There exists a vast amount of algorithms aiming to select only the relevant features

The above is a classifier model on which LingPipe is based.

For this, the LingPipe code is given below-

```
import com.aliasi.classify.ConditionalClassification;
import com.aliasi.classify.LMClassifier;
import com.aliasi.util.AbstractExternalizable;
public class SentimentClassifier {

    String[] categories;
    LMClassifier lmc;

    public SentimentClassifier() {

        try {
            lmc= (LMClassifier) AbstractExternalizable.readObject(new
File("path for the classifier"));
            categories = lmc.categories();
        }
        catch (ClassNotFoundException e) {
            e.printStackTrace();
        }
        catch (IOException e) {
            e.printStackTrace();
        }
    }

    public String classify(String text) {
        ConditionalClassification classification = lmc.classify(text);
        return classification.bestCategory();
    }
}
```

4.3 To test for live tweets

The system for getting live feed of tweets and using the 2.5 lakh+ dataset as training data, and then testing whether at the particular instant, which party is trending more positively or negatively. The GUI for this is shown below-

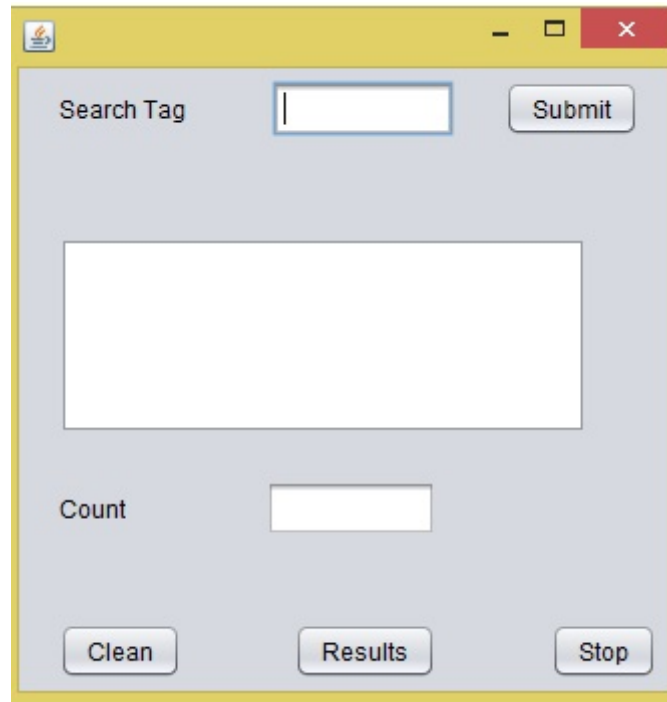


Fig 4.5 GUI for getting live test data

CHAPTER 5

RESULTS AND EVALUATIONS

5.1 Analysis of collected tweets

After collecting 2.5 lakh tweets over a period of time, using our search query terms, “bjp”, “modi”, “cong”, “gandhi”, “aap”, “kejriwal”, we have found out that most discussed party is AAP over twitter.com with more than 1.2 lakh tweets. On the 2nd position came BJP, with little less than 1 lakh tweets being collected. Congress was least discussed or trended over twitter with about 50,000 tweets.

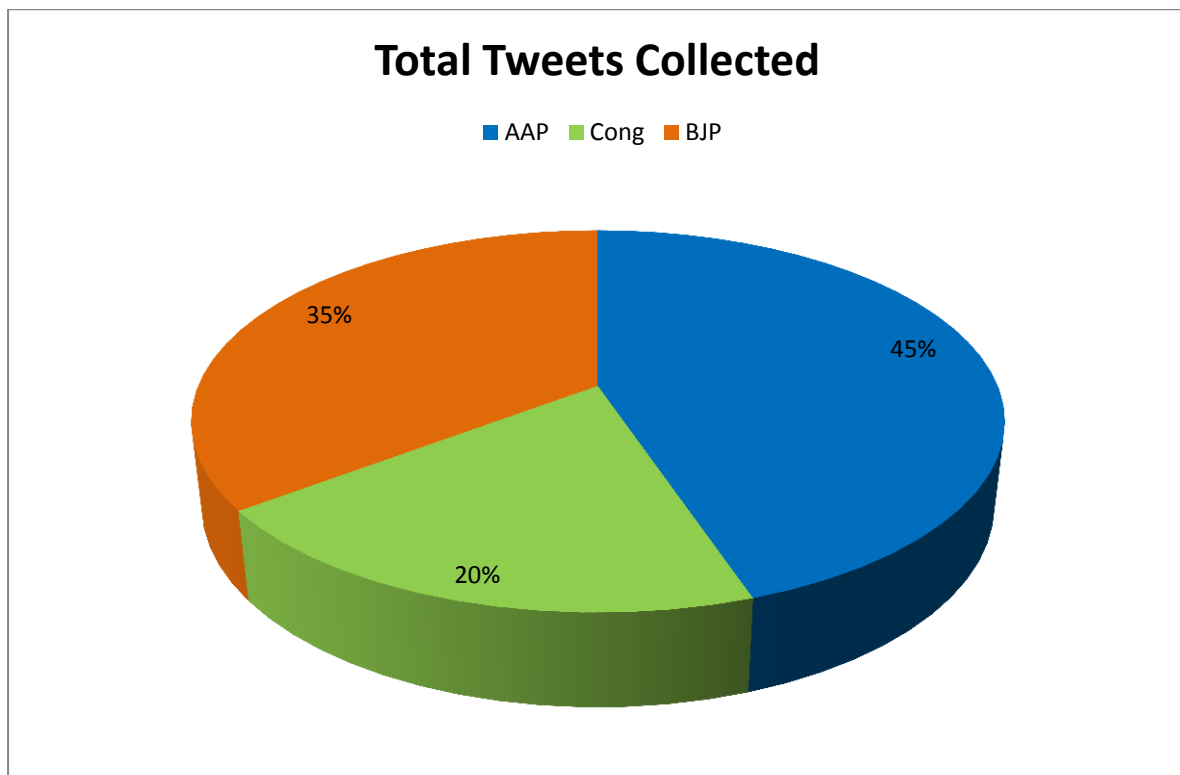


Fig 5.1 Total tweets Collected

5.2 Analysis of positive or negative tweets

The analysis of the positive and negative tweets as generated from LingPipe are given below-

5.2.1 BJP

The analysis of BJP with the keywords used as “bjp” and “modi”. This analysis has come largely in-line with what the opinion polls have predicted and were conducted during the period when the data collection was taking place.

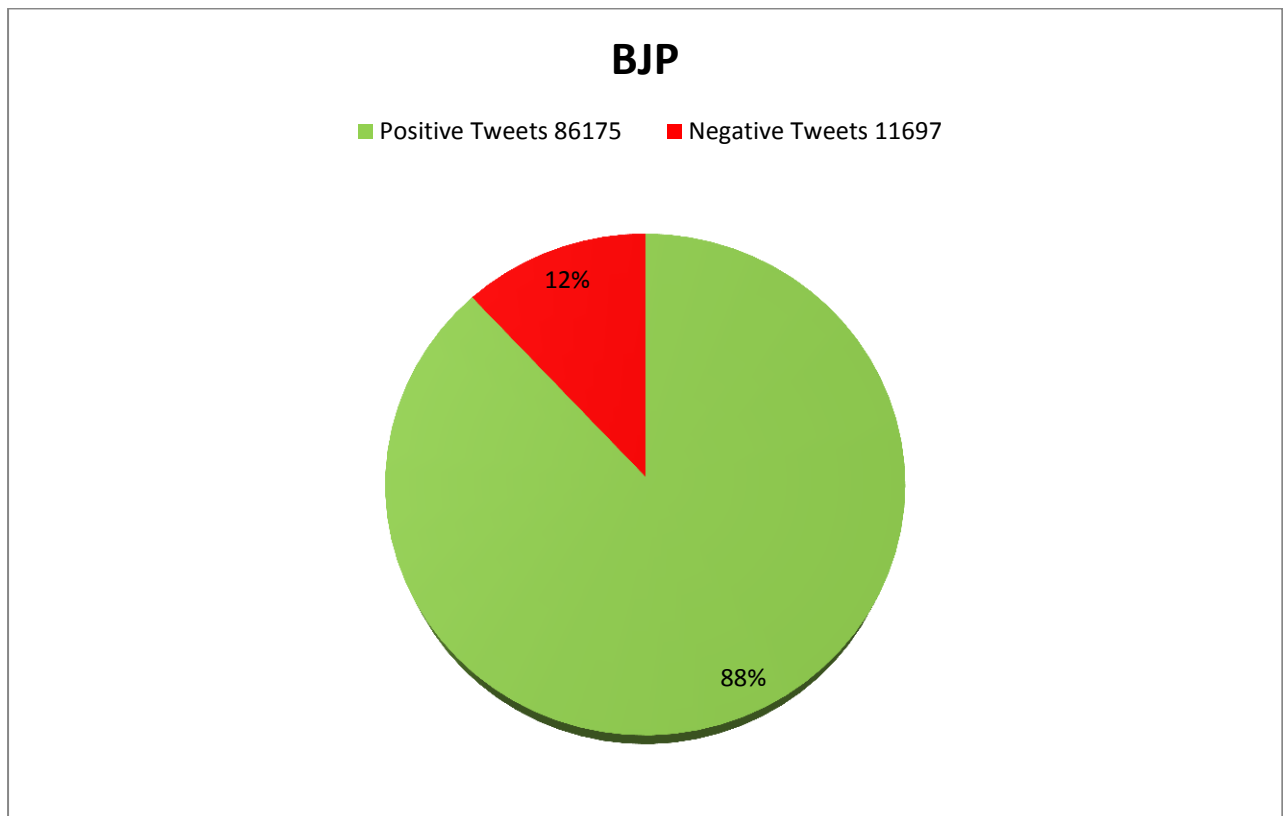


Fig 5.2 Figure showing result of sentiment towards BJP

5.2.2 Congress

From a total of 50k tweets, the analysis shows that about 18% are in favour of Cong, and a huge amount of about 82% tweets are in negative.

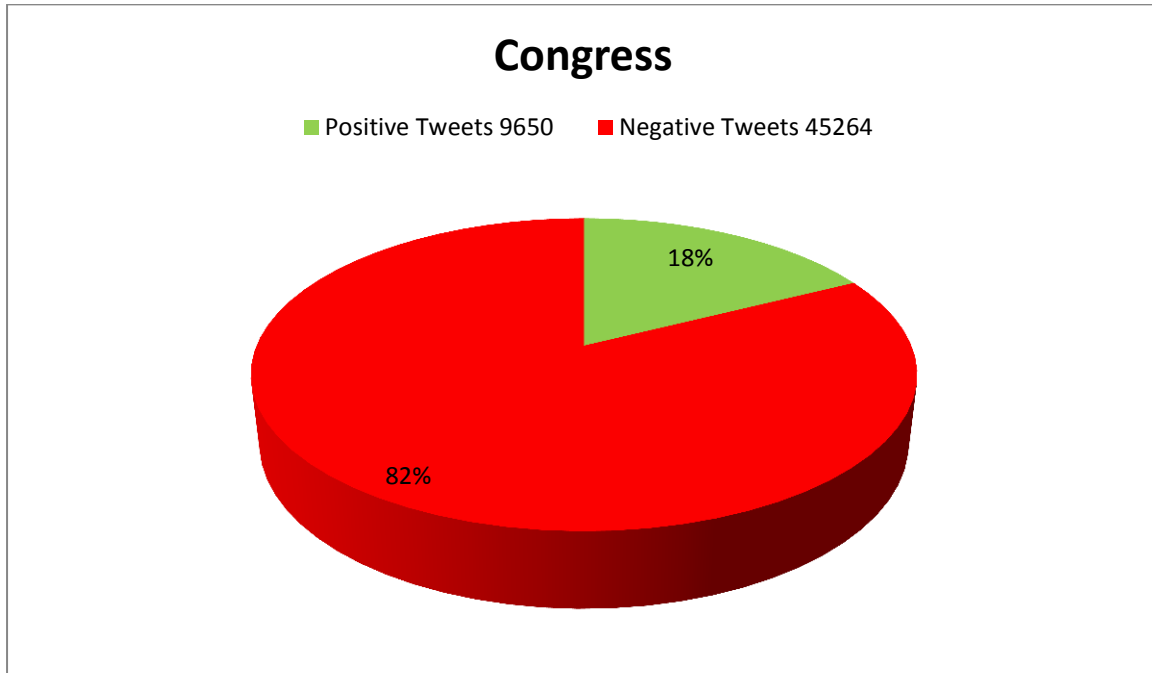


Fig 5.3 Figure showing result of sentiment towards Congress

5.2.3 AAP

From a total of 1.2 lakh tweets, the analysis shows that about 40% are in favour of AAP, and 60% tweets are in negative.

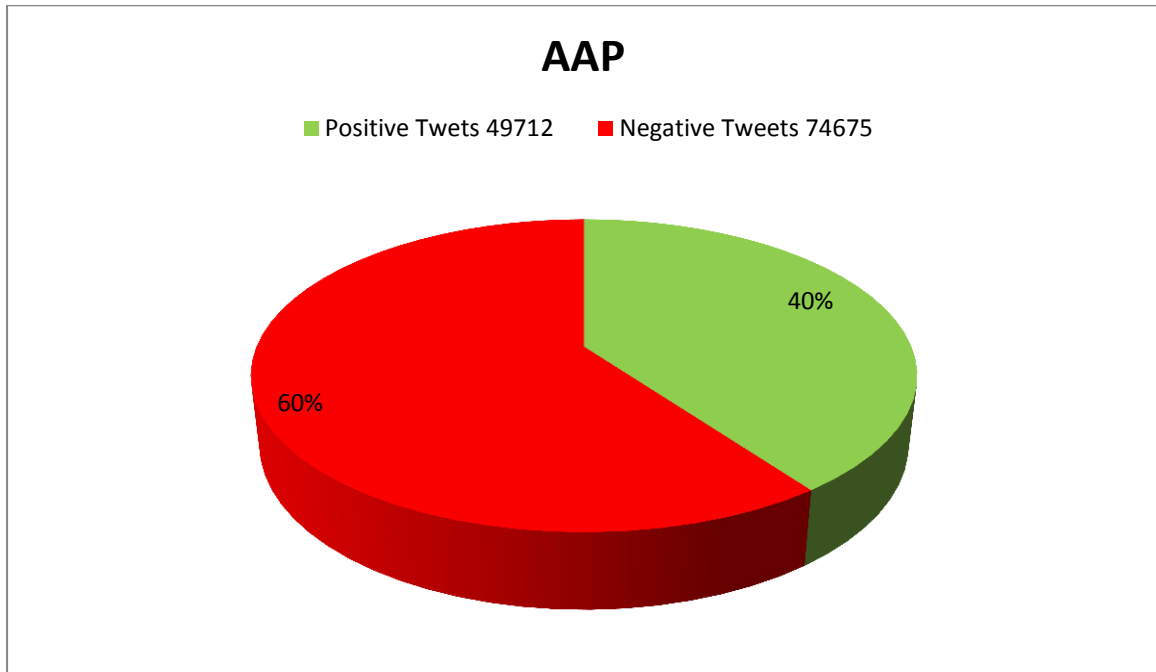


Fig 5.4 Figure showing result of sentiment towards AAP

CHAPTER 6

CONCLUSION AND FUTURE WORK

This report summarizes a way to categorize twitter messages or tweets into various sentiments and focus here is only on positive and negative. There could be very positive, very negative and neutral also. This report focused on 14th Indian General Elections, 2014 so as to study about the support towards the various big political parties in the elections. The focus to get tweets was on BJP, Congress and the newbie AAP. AAP is said to be a party born because of the social media and the present form is greatly influenced by the social media. Also it is widely reported that political parties have established special cells in their workforce to manage their social media image.

So, we collected huge amount of tweets so that better analysis could be done and we can get a big picture of which political party is getting maximum support on twitter.

The analysis has shown us that AAP trended the most on twitter followed by BJP and Cong. The analysis of positive and negative among them showed a different picture, i.e which was most trended doesn't means that it had more supporters.

In our analysis, BJP has got 82% positive tweets whereas 18% negative tweets about it. AAP ratio stood at 60% in negative and 40% in positive and Congress was very badly perceived with 82% in negative.

This analysis doesn't imply that the resulting seats outcome of the result would be based on this as in India, there is 16% reach of internet in India and of them only 35% talk about politics on social media.

So our analysis show that BJP is the most favoured party on twitter, i.e maximum positive support is there for them on twitter.

Also now using this 2.5 lakh+ dataset as training data, we have created an application, to test for live tweets, that is at a particular instant which party or which political person is trending positive or negative.

Using this procedure in future, we can do the same analysis for any phenomenon or product by using a specific search query and then first collecting few tweets for training we can detect the sentiment towards the product.

References

1. R. L. Brown. 1980. *The pragmatics of verbal irony*. In R. W. Shuy and A. Snukal, editors, *Language use and the uses of language*, pages 111–127. Georgetown University Press.
2. Sitaram Asur and Bernardo A. Huberman. *Predicting the future with social media*. Mar 2010.
3. Avrim L. Blum and Pat Langley. *Selection of relevant features and examples in machine learning*. *Artificial Intelligence*, 97:245{271, 1997}.
4. A. Go, R. Bhayani, and L. Huang. *Twitter sentiment classification using distant supervision*. 2009.
5. Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. *Social networks that matter: Twitter under the microscope*. CoRR, abs/0812.1045, 2008.
6. Dunlap, J.C., & Lowenthal, PR. (2010). *Tweeting the night away: Using Twitter to enhance social presence*. *Journal of Information Systems Education*, 20, 129-135.