DETECTING REASONS OF NEGATIVE PUBLIC MOOD USING TWITTER

A

Dissertation

Submitted in partial fulfillment of the

requirements for the award of the degree of

MASTER OF TECHNOLOGY

In

ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWORK

By

Mayank Sah

Roll No.: R102213005

Under the guidance of

Ms. Poonam Kainthura

Assistant Professor, CIT, UPES Dehradun

Department of Computer Science & Engineering

Centre for Information Technology

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, UK

May – 2015

CANDIDATE'S DECLARATION

I hereby certify that the dissertation entitled **"DETECTING REASONS OF NEGATIVE PUBLIC MOOD USING TWITTER"** in partial fulfilment of the requirements for the award of the Degree of MASTER OF TECHNOLOGY In ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWORK and submitted to the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my work carried out during a period from **January**, **2015** to **May**, **2015** under the supervision of **Ms. Poonam Kainthura, Assistant Professor, CIT, UPES Dehradun**.

The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other University.

(Mayank Sah)
Roll No.R102213005

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: _____

(Ms. Poonam Kainthura)
Dissertation Guide

Dr. Amit Agarwal
Program Head – M.Tech (AI & ANN)
Center for Information Technology
University of Petroleum & Energy Studies
Dehradun – 248 001 (Uttarakhand)

ACKNOWLEDGEMENT

I wish to express my deep gratitude to my guide **Ms. Poonam Kainthura**, for all advice, encouragement and constant support she has given me throughout my dissertation work. This work would not have been possible without her support and valuable suggestions.

I sincerely thank my respected Program Head of the Department, **Dr. Amit Agarwal**, for his great support in doing my dissertation in **Sentiment Analysis** at **CIT**.

I am also grateful to **Dr. Manish Prateek, Associate Dean** and **Dr. Kamal Bansal, Dean CoES, UPES** for giving me the necessary facilities to carry out my dissertation work successfully.

I also express my gratitude to **Prof. N.L.Sarda, Prof. Jitendra Shah,** and **Prof. M.Edward Barowski** for their constant guidance and constructive criticism.

I am very grateful to my respected class coordinator **Mr. Vishal Koushik**, for showing a very keen interest in my work and motivating me to expand the horizons of the work.

I would like to thank all my **friends** for their help and constructive criticism during my work. Finally I have no words to express my sincere gratitude to my **parents** who have shown me this world and for every support they have given me.

Name **Mayank Sah**

Roll No. **R102213005**

ABSTRACT

The scope of social media is ever growing. In today's world it is one entity which attracts us to an extent where one regularly generates some relevant social data. Whether this data be a simple like, or a tag of photograph even some recommendations. Thus we all generate social data. But the question arises what happens of this data, and does this data has any significance at all. Thus to answer these questions, we have designed a project which takes the social data and analyzes the data for its importance. The importance is a sentiment value which help predict public sentiment about any topic. The analyzer system uses a lexicon based approach to detect the public sentiment and then uses a variation in Bayes classifier algorithm to extract the keywords causing the sentiment value. Thus not only giving us an approach to use social data but also give us a reason for the mood. The testing of sentiment analysis has been done on 1000 tweets, containing the keyword disaster. This was then analyzed using the algorithm and found that it is a negative sentiment word. And then using a probabilistic measure some words causing these negativities were identified. Thus in the vibrant realm of social media this project works as a minute system to give us an informative output from a very complex and ambiguous dataset.

CONTENTS

LIST OF FIGURES

Figure **Page No**

CHAPTER-1
INTRODUCTION

1.1 History of social media:

Human being is a social animal. We are social as we feel the urge to communicate with each other rigorously for many purposes like enjoyment and knowledge gain. The roots of social media are rooted beyond our imagination in our lives. Although with the advent of applications like facebook, twitter and linkedin, we feel that this is a new age mechanism.

In early years of communication hand dropped letters were used as a communication media. Postal services dates back as far as 550 B.C and since then the postal service has been constantly a part of our social lives. In 1792, the telegraph service came into picture. The telegraph drastically changed the speed of communication, these were short messages, carried through a communication link. Thus the telegraph became a revolutionary way to interact and share information. Then after the hang of telegraph service started dropping a new service named as pneumatic post came into our lives, developed in 1865, this new service paved a new and efficient way for delivering letters quickly between recipients. The pneumatic post uses a set of information capsules to be transferred using well-grounded pressured air tubes from one place to another.[1]

The last decade of 1800s gave us two major boons, which revolutionized the entire communication system. These two phenomenal discoveries were telephone in 1890 and radio in 1891. Along time ago, when we were not netsavy, and the internet had not stretched its wings in each and every paradigm of our daily lives, from study to entertainment, a new kid on the block was coming to become a giant. This kid started its journey from the research labs and then graduated to be used by the army and scientific community and started to invade our live. This new kid on the block was computer network, and the time was early 1950s.[1] Computer networking started in its early years as a tool for carrying out military based operations. But as time flew by, computer networking invaded every paradigm of our lives starting with a simple telecommunication network and later on to much sophisticated activities. In the 1970s, with the advent of IT revolution across the globe, the transitioning of computer network use rocketed sky high.[2]

Late 70's and early 80's were moreover dedicated to type writer , and computers were a rare commodity destined to be found only in research labs or with prestigious colleges. The use of computers was not very accessible and very hard to learn, and thus making its usefulness very limited, because for using a computer a person had to be well versed in programming languages, which at that time was a much closed set of people. To make it worse the entire hang of sitting in front of a computer was a very boring and lonesome experience. Thus we can really understand that computer network was only used by a set of really ardent and great computer enthusiasts. Most of the people who grew fonder of this system named as computer and the topology which was named as a computer network were a type of outcasts, who in general did not mix well with others. But it was only due to these outcasts and socially uncomfortable people that the social web we see today blossomed. The people once considered as nerds and geeks are now a social enigma, visionary and leaders. The social networking thus in 2015 has changed to be a high level of computer networking where each node is a person and all its passing edges represents its social status. The very first of such social network sites were BBS and AOL.[2]

1.1.1 BBS, AOL and CompuServe: The Infant Years

Bulletin Board System (BBS) started out as a meeting place for a set of computer enthusiasts, where all of them met online and share with each other their chunks of codes and games using a central system to monitor the entire thing. Later as BBS started becoming popular they started using chat services and minor messages of communication using a telephone line and a modem. BBSes became a very popular medium of communication among those who now started mixing up with each other over internet in a business related environment and also helping one another in solving complex computer problems. In other words it started functioning as a multifunctional, and a well facilitated laboratory for innovators. But as BBSers were using a telecommunication link, thus out of area call charges were applied for their communication, which used to cost more. Thus the local communication increased a lot, and a substantial group of societies started communicating over BBS, and valla the antisocial had now become social.[2]

The BBS was not just simple fun and games and everything was not all hunkadory, the slow speed of connection and the one dimensionality of the service was not at all the seventh heaven, but still this innovation was much appreciated till late 80s and early 90s. [2] Some services such as Fidonet ensured functioning of BBS even in the internet era when more advanced

applications started emerging in the world. Fidonet combined a set of BBS users together and made interactions a bit interesting graphically. But every good or bad thing must come to an end, and so did BBS. The mere chatting service and technical complexities rigorously declined the use of BBS and other improvements to BBS started replacing the network, which was once colossal. Out of many of these improvements, CompuServe emerged as a leader. CompuServe was started in 1970s for a mainframe type communication scenario, for research purpose, but gradually came into public domain, to replace the king and rule the realm of social media. [2]

CompuServe allowed its users to chat as in BBS but on top of it also laid the first ground stones for public forum. All the users of CompuServe could post a problem and all the members can give a feedback on this problem. In this process CompuServe also established a vague trend of distributed computing. This application also generated a new and profound method of communication, the e-mail. This technology has become a very important part of our lives and then it was for the very first time incorporated in an application.

But the progress came to its full exuberance with the introduction of AOL (America Online). It is considered the first full-fledged social networking site. AOL is fondly called as “The internet before internet”. The members in AOL were well equipped to create communities and access profiles of other members as well. This new feature of profile and accessibility of this profile caught the market like fire in a jungle. The fascination of social media had now started flying.

By the time AOL had spread its wings, the flower of internet had also starting blooming, and it was mid-90s. Giants like yahoo and amazon started in this era. Apple encouraged people to buy computers and brought the computer technology from specific people and communities to every household. Now amazon had started selling books online, yahoo had started capturing its own market and the social media we know today started shaping on.

1.1.2 The Internet Boom (Adolescence Age):

The ideology of social networking now started changing in this era. Instead of looking for anyone with whom you can connect, people started searching for people they know in real life and connect with them in the virtual world also. One such game shifter was a site Classmates.com. This website popularized on the back of passing school kids, who want to be connected with their classmates even after they have graduated. Users could create profiles, and search for long lost

batch-mates and share their journey and relive old times. This site even today has a 57 million odd user accounts registered.

One such attempt did not live long enough to blossom was SixDegrees.com. It started in the year 1997, based on the theory of Hollywood actor Kevin Bacon that no two persons are separated from one another by not more than 6 degrees.[2] It was one of the first sites to enable users to create profiles, communities and arrange their communities. The users could surf other profiles as well. But this site disclosed a dark side of social media known to be as "Spamming". The spamming on this site disgusted the users and sixdegrees.com died with the same speed with which it had emerged.

With the downfall of sixdegrees and the competition arising in social web, again altered the thinking, and in came the era of demographic-oriented networking sites. AsianAvenue.com, BlackPlanet.com and MiGente.com are some of these area oriented social websites. Out of these websites MiGente.com is still working with nearly 8 million users at hand.

1.1.3 Friendster, LinkedIn, MySpace and Facebook: The Biz Grows Up

As we stepped into the 21st century, the realm of social media expanded to greater heights. In 2002 a social networking site named Friendster was launched. It was much similar with sixdegrees.com but instead of searching for six mutually known friends it started creating circles of friends and the concept of mutual friends stepped in. [2] On basis of these mutual friends the site would share a common interest among these friends and hence grew the social bonding among the friend circle. Each friend circle would be entertained on basis of their mutual interests, thus creating ample lot of opportunities to discover their feelings and interests.

On basis of these mutual interests Friendster emerged as one of the first dating sites in the world. It had more than 3 million registered users and the revenue generation was astonishing. [2] Friendster also promoted the idea of gaming in communities and thus propelled in all aspects of social aspects. But with time it started becoming boring and slowly regressed. Now it is only available as an online gaming service.

Just after the emergence of Friendster, in came an ace idea of social media. LinkedIn was introduced in the year 2003. [2] LinkedIn was a much more organized and focused. It focused on the business people, looking ahead to make a career. LinkedIn provides a set of professionals to communicate with each other and update their business experience. Various communities

accessible in LinkedIn helps the freshers coming into marked to land up a good job profile by getting to know the business leaders and deciding the way they can achieve a good carrier. It also enables people from same class of markets to come together and solve their business problems. Each contact in linkedIn is known as a connection. People can update their skills and profiles to make them hirable. Linked has more than 297 million registered users and still growing. [2]
Another prodigy launched in 2003 was myspace. MySpace was once at the pinnacle of social webspace, as it supported demographic music and videos in a feature rich interface. Youngsters took to myspace instantly and also served as a platform for budding talents to publicize their skills in videos. However once touching the skies myspace has also suffered with the tough competition and now is just bounded to promoting videos and music files among young musicians and upcoming bands.

To dethrone the king myspace, in came a worthy opponent “Facebook”. Facebook started as a university project in Harvards. It started off as a publicity ground for students to publicize their achievements and products to other students of the university. Facebook was launched in 2004 as an internal website at harwards, and worked for 2 years as an internal social group. In year 2006 facebook became public and took the world with its innovative style and great interfacing. Various facilities like sharing videos, photographs, opinions on social platforms lifted the social media game multiple notches up instantly. From the time it was made open for public facebook started attracting huge investments, and the realm of social media grew richer by the day.
Facebook currently boasts of more than 1.3 billion registered active users and is the most used social networking site across the world. [2] The reason for the much appreciated facebook is the multitude of features and the comfortable accessing of these features. It enables people to be much more open about themselves, the world around them. It also works as a very good medium for promotions of individual’s skills or even products. Thus due to this large promotional ground facebook is a sae heaven for investors and continue to attract investors.

The break to fame for facebook is not a fluke but rather the brilliance of founder Mark Zuckerberg. It is not only the most innovative site on social network which keeps updating but also encourages the change. Facebook was launched with an open platform, where third party developers can create their own applications and add it to work with facebook. This open API interface encouraged the app developers and game developers to provide variety of applications for facebook users. This promotion to app building became such a big success that facebook had

to create a facebook app store to keep all the applications available to users. Inspired by this innovative idea Twitter also created its open APIs and enjoyed the same if not more success than facebook.

One point which became an identity for facebook and contributed massively making facebook popular, and that is it's like button. It gave the users a simple medium to appreciate everything around them. The introduction of like proved to be a masterstroke for facebook. It even broke the boundaries, and went out and became popular as much as that in every phase in cyberspace we see a like button. Similarly tweet worked in favor of twitter where everything can be tweeted. Realizing the massive popularity and business growth in social media, Google also jumped into this field with the introduction of google+ in 2007. Google+ became popular because of the complete experience of social interaction. Its hangout feature was accepted with open hearts as this would enable users to start a video chat any time with a set of friends and also share the views in a well dedicated community circle.

The success of google+ also owed a lot to the gmail service over which google+ is engraved. As people who are in gmail become a part of google+ as well. So you don't have to visit two accounts but rather your one account works for all your interactions. Google+ inducted 25 million users within a month of its release. The competition still grows and the innovations are keeping both facebook and google+ on their toes.

1.1.4 The Multi-platformed Self: The Rise of Mobile

With the advancements in mobile industry within last couple of years, and introduction of smartphones and tablets have altered the entire social networking experience. We are no longer bound to sit in front of a computer screen to enjoy our social circle, but now it has become our every time companion. We are now round the clock connected with our near and dear ones on social front using social media apps. Thus mobiles have not only taken social media to our pockets and hands but to our hearts.

Mobile computing is now a detailed field of study. With the advent of android and windows phones social media platforms had to adjust to the change, and adapted they have. Newest members of this social media family are applications as snapchat and instagram. Instagram alone has generated a magnificent 20 billion images since the time it released in October 2010. [2]

Instagram has made each and every moment in a day a probable digital memory, with the good picture value and emotional swings faced in a day by usrs.

The approach of mobile devices in accessing social media is completely different from the desktop devices. The mobile experience of social media has reduced the expression of words but promotes the interaction in images. Public sharing and private sharing of images has overwhelmed the web and now the focus is shifting on location based applications like tinder, to augmented reality applications like Foursquare, private sharing of images like snapchat and public sharing of images like instagram. Thus mobiles have changed the complete outlook towards social media.
Sites like twitter encourage people to publicize their opinion in short words and share the happenings continuously. The emergence of twitter has given a new dimension to social issues tackling and how to influence public mood.

1.2 History of Machine Learning:

Machine Learning is a subfield of artificial intelligence. The development of artificial intelligence consists of the continuous development of machine learning. The basic concept of artificial intelligence is to enable machine or computer intelligent enough to make decisions autonomously. The year 1957 witnessed the induction of perceptron model, which stated the value of perceiving from environment. [3] The perceptron model also paved the way for inducting a learning mechanism for machines. And with the sheer brilliance of the concept there was a positive vibe regarding artificial intelligence and amalgamation of machine learning in its future till 1960s. [3] But the perceptron model had some limitations in expressing complex functions as pointed by Marvin Minsky, thus demotivating researchers to pursue machine learning as the next big thing for the entire next decade.

The demotivation was so intense that in 1970s the field of machine learning was almost extinguished, with only a couple of enthusiasts still willing to invest their time and energy in exploring machine learning possibilities. This was the decade for expert systems, and the progress in the entire field of artificial intelligence was measured in the advancement in the field of expert systems. 1980s were a resurrection for machine learning paradigm with the introduction of Decision trees and artificial neural network. The redemption for machine learning came in mid 1980s when decision tree model came into picture, and was distributed in form of complete package. This decision tree model is highly simple for human eye to understand, and makes the

entire explanation and procedure very simple and exciting. The realization of a multi-layer neural network, saw the limitation of perceptron model removed, with enough number of hidden layers designed for a problem.

The applications of decision trees and artificial neural network are very varied, and effect multiple facets of computer usage, for example: loan approval and fraud detection and portfolio management. These machine learning techniques paved the way for an autonomous learning mechanism and a complete automation approach. Figure 1.1[3] shows the history of machine learning and the transition of paradigms over the years.

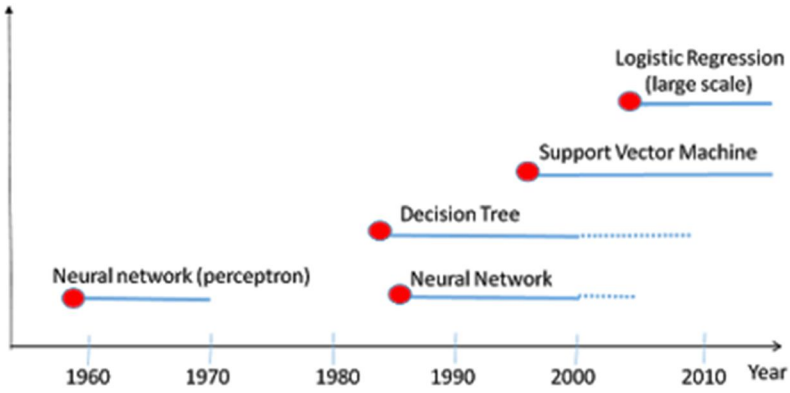

Fig1.1: Brief history of machine learning strategies

With the advent of internet in 1990s, the growth of machine learning grew exponentially as the data could be accessed easily and the boom in data made the training job easier in machine learning, as sufficient samples could be arranged. With the growth in use of internet, management of data and internet started becoming tougher by the day, thus the need of automation in data management rose. In 1995 the concept of support vector machine SVM was published, and since then has been widely accepted for making machine learn. [3]

On the footsteps of 21st century, logistic and statistical regression techniques blossomed for solving machine learning problems. And in the first decade of the new millennium machine learning algorithms have come a full circle with advancement being made in every part of the algorithms and the increase of machine learning use in developing intelligent systems. After year 2000, Logistic regression was rediscovered and re-designed for large scale machine learning

problems.[3] The new methods were received and analyzed rigorously, out of which Bayesian classifiers, and maximum entropy classifiers seem to be making the learning much more effective.

1.3 Mining the social web:

The continuous growth of social media has made the cyberspace a whole new world in its own. The lively beings of this new virtual world are we the people, who use various applications over the internet in our daily lives. Social media has given a different type of power in our hands. Each one of us is now showing our opinions and our interests in various social networking sites. These opinions can be used in multiple capacities, so as to detect mood of public regarding an event. It can also enable us to work as citizen journalists, where we report each and every incident which takes place in our vicinity. These citizen journalists can give us a more real and precise account of the activities taking place.[4] Thus processing this newsfeed requires mining activity. Hence this word social media mining came into picture, which deals with implementing various techniques and methodology to extract some chunk of knowledge from the social networking websites. This extracted chunk of knowledge might be an opinion, a decision or just a summary of events occurred. The mining of the social networking sites also enables us to understand the changing human psyche and views on different topics and problems surrounding us all in real life events. This social media might also be used as a citizen charter where people decide on the policies and laws to be passed by government and add a fifth pillar to democracy. But analyzing this data from social networking sites or simply called social media is not at all an easy task.[4] The sheer bulk of data travelling in social media makes it hard to process and extract information from social media. Plus the various existing relationships as friends, circle, followers and followees makes it even tougher to evaluate, and thus calls for some new refined methods for information extraction. [4]

Social media analysis can be a very good in identifying the basic functionalities of group formation among individuals and also analyze the trending topics among individuals to make a much better profitable business environment for economic and social growth of any nation. Social media mining can be very effective and economical in terms of political relation maintenance and monitoring public mood over government policies, and the topics of interest for entire nation.

Microblogging site twitter can be a very good source of data to be analyzed for extraction of information. Twitter data contains a lot of meta-data associated with the objective data to give

us a perfect medium for applying social media mining, effectively. The content is limited by the number of words used thus noise removal also becomes easier, hence giving us a perfect ground for social media mining.

Fig1.2 [5] displays the various methods for analyzing social media

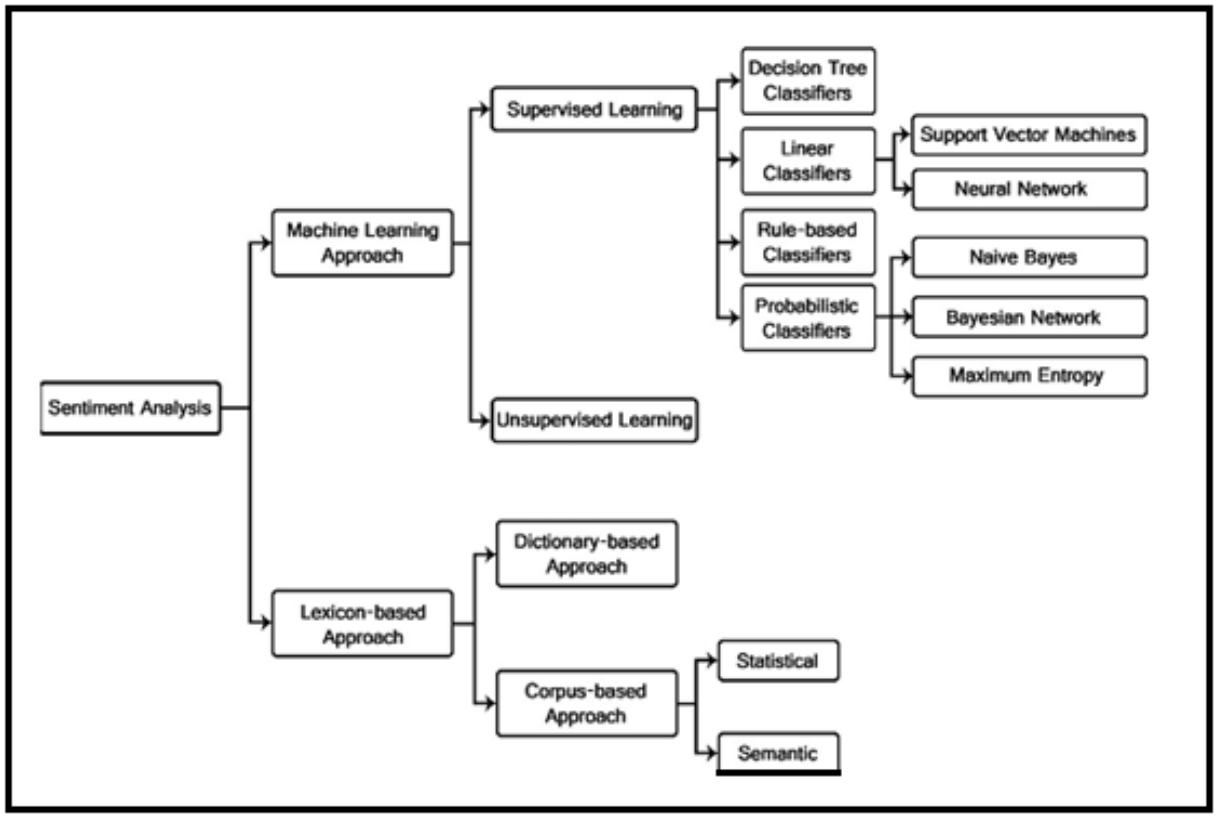

Fig1.2: Methods for mining the social media

Microblogs are an opportunity for scavenging critical information such as sentiments.[6] This information can be used to detect rapidly the sentiment of the crowd towards crises or disasters. It can be used as an effective tool to inform humanitarian efforts, and improve the ways in which informative messages are crafted for the crowd regarding an event. Unique characteristics of microblogs (lack of context, use of jargon etc) in Tweets expressed by a message-sharing social network during a disaster response require special handling to identify sentiment. We present a systematic evaluation of approaches to accurately and precisely identify sentiment in these Tweets.

Microblogging sites are used by individuals to post real time messages. [6] These messages might be about any event occurring I their vicinity or any social tragedy or any view regarding any event taking places in their lives. It gives the users a platform to publish complain as well as their

admiration for some work or some people, giving us an excellent hunting ground for scavenging information.

One part of social media mining is sentiment analysis. Where using the social data we predict the sentiment or mood of people regarding any keyword or topic. [6] There can be three basic categories of sentiment under which the data might fall. These three categories are namely- positive, negative and neutral.

Avery popular social networking site that can be used for analyzing public mood is twitter. Twitter contains short messages named as tweets with a maximum word length of 140 characters per tweet, giving us a less noisy data set, as we have to analyze less number of words per tweet. Twitter uses some unique identity symbols, in order to define the data.

- **Emoticons:** Emoticons are used to symbolically define user moods. Various facial expressions put in the tweet to establish an emotion. [6]
- **Target:** Target is the entity or person about whom the tweet has been posted for. The target is defined by a "@" symbol, followed by the name. [6]
- **Hashtags:** Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets. [6]

1.4Purpose

Our main purpose is to use twitter as a source of data and classify the sentiment from these tweets, in order to grab an essence of public mood regarding any activity or any event. With the boom in social media equipment, with mobile computing coming into picture and omnipotence of twitter apps in mobile phones, each person has become a citizen journalist having the power to report everything happening around him. Thus all important topics can be searched for using various twitter APIs. Twitter approximately generates 170 million tweets every day. [6] Thus we can imagine the extent of reach of social issues from twitter. This twitter data is then classified into three sentiments- positive, negative and neutral, giving us an entire sentiment regarding any activity taking place. We use a lexical approach for deciding the sentiment. Our objective extends in finding out the reason, if there is a negative sentiment surrounding the event.

After the tweets have been searched for and are put together we intend to get a reason for the negative. For getting the reason of this negative sentiment, we use machine learning approach, with a statistical naïve Bayes classifier, to check for the various reasons for any negative sentiment

surrounding an event. The words are identified using a keyword extraction strategy to treat the keywords as a reason for negative feedback. There might be a single reason for negative feed or might be multiple reasons behind something negative. This project intends to find maximum reasons to any negativity in the public mood, so as enable automatic evaluation of sentiment and reason clarification.

1.5 Outline of the Dissertation

This thesis is organized into six main chapters. Chapter 2 introduces the previous work done in sentiment analysis and machine learning under the heading related work. It is a basic literature survey of the various papers and books referenced in order to carry out this project. Chapter 3 of the thesis explains basic challenges in mining the social web space. Chapter 4 describes the background study for understanding this project. It explains various basic terms like sentiment analysis, machine learning and other prerequisites to make a good understanding of the project. Chapter 5 enlists the various requirements for carrying out the project. The requirements part is subcategorized into hardware requirements, software requirement and package requirements. Chapter 6 explains the methodology used to achieve our objective. It also explains the mathematical model and analyzer algorithm to be implemented. In chapter 8 the implementation pseudo codes are mentioned, with detailed algorithms. And finally Chapter 8 constitutes results achieved from the project. Chapter 9 concludes the thesis and elaborates the scope of effect in real situations.

CHAPTER-2
LITERATURE SURVEY

2.1 Social Media Mining and Sentiment Analysis:

Hogenboom et al [7] explains the utility of emoticons in text. The paper deals with analyzing how various emoticons can have meaning in multitude. They discuss about designing a new model for analysis of these emoticons, which may lead to a more effective sentiment analysis. To improve the efficiency of lexicon based sentiment classification method the paper emphasizing in creating a manually annotated emoticon dictionary to give an accurate meaning while analyzing. For the testing of this new approach 2080 dutch tweets and forum messages, which contain emoticons. After testing the tweets sentence wise, they concluded that wherever emoticons are used, it reflect the sentiment more accurately than mere text. [7]

The research paper written by Choi et al.[8] deals with the syntactic and semantic linking of keywords with their sentiment. Specifically, one word may have different sentiment value in different domains. Thus to tackle this problem of variable sentiments for same keywords on different domains a new framework has been proposed. This new framework prepares a corpus of the entire document at hand and then semantically analyze the document to give us a measure of the domain of the context. When the domain has been clearly identified the features of the corpus are developed. Then a bootstrap algorithm is used to extract sentiment clues from the document. Sentiment clues are the basic probabilities of sentiment in different use cases or domains. For experimental purpose four domains are specified, namely business, international event, environment and politics and a domain wise sentiment analysis was done. As a result it was observed that the sentiment values are syntactically attached to a corpus and same word might give different sentiment value in different domains. [8]

The research undertaken by Maks and Vossen [9] expresses the need to dig deeper into sentiment analysis by extracting a sentiment using verb phrases. This is an advancement on the analysis as no single keyword is outlined to generate sentiment score, rather the verbs in a sentence and its usage plays a part in the sentiment classification. All verbs are kept in seven semantic categories and three states are kept to deliver emotion value. These three states are then recursively aligned with the seven semantic categories to realize the subjective bonding between multiple

attitudes, as where and how the verb is being used. As a result of this study a list of 580 verb use cases are prepared to give a sentiment score, which can be used automatically to detect deeper sentiment value by using simple matching algorithms. [9]

Another paper written by Hare et al. [10] deals with the issue of analyzing financial blogs in order to use the sentiment value in stock market prediction using public sentiment shift in the market. The topic deals with the importance of financial blogs and how these blogs shift the mood of public. For analyzing the sentiment from the blogs, a corpus based approach is proposed, where the corpus is created in six levels of creation and then analyzed over three main criterions, namely annotation statistics, inner annotator agreement and topic relevance. The sentiment values are generated and the shift in public mood is observed monitoring the inner annotator shift.

An idea of a weak supervised sentiment analysis by adding historical sentiment value to any topic and generate sentiment labels based on this historical data. Two approaches are used in order to label the historical sentiment value was proposed by He. [11] He uses two approaches where one approach is using LDA – DP, where the dirichlet prior are modified prior to the topic word distribution, by using a Bayesian classifier to do so. The other approach is adding a preference value to each topic term and generating labels using general expectancy to follow the sentiment label. This approach is said to be LDA _ GE. The experiments were conducted over movie reviews, books, and product reviews for sentiment labeling.

The book written by Russell [12] deals with the twitter API and how to work with tweets using python framework. It also explains the graph structure of facebook and how to utilize likes and comments in facebook. It also seeks data mining, analysis, and visualization techniques to explore data and answer the following representative questions:

- How to determine the entire friend circle of individuals?
- How often people interact in these friend circles?
- Which of the circles generate good opinion value about a ceratin keyword or activity?
- Does geographical boundaries play role in online friend circle creation?

This book also extensively talks about the graph structures being followed by the analysis of these graphs to analyze the web of social media.

Theil et al [13] furnishes the result of a market survey by analyzing the sentiment and the process of sentiment creators or influencers, who try and influence public mood using blogs and review sites. For achieving this goal he frequent reviewers and influencers were monitored online

and the sentiment they publish about any entity. With rigorous training it was observed that people giving a negative sentiment n general do not influence market in that capacity, as the critics are affecting less product. This white paper also introduces the basis of a sentiment scorecards. These score cards are very good in keeping an overview of the entire analysis.

In the book written by Zafarani et al. [14] the very basic architectures related to mining the social web are discussed descriptively. Starting from the basic network topologies and the type of networks used in order to create a social cyberspace, to the advanced algorithms used to extract knowledge out of this very unstructured and noisy data source. This book also explains the various aspects of mining social web with the right ingredients. The extensive discussion about how to inculcate a multidisciplinary approach including machine learning, computer network, data mining, statistical techniques in order to make social media a vibrant data source. The later part of the book offers an insight to group and community formation in cyber space. This part is followed by the applications of mining media from social networking websites, and also the challenges faced in order to do so. The book concludes with the behavior and pattern analysis of individuals by using the mining frameworks.

In his research paper Liu [15] starts with the basic introduction to the field of sentiment analysis and describes some technical bottlenecks faced during the analysis. It is a very good review of the domains that can be influenced using sentiment analysis as a tool, and thus on basis of the varied scope of implementation, terms sentiment analysis as a multi domain multifaceted problem. These problems or domain are further explained as subdomains and the depth of varied influence capacitated by the sentiment analysis module in our daily business lives. The writer has graciously shared his thoughts on the past, present and future of sentiment analysis and termed the ever growing field as a force to be reckon with in the future.

Jin [16] has very nicely explained the power of **'**Like' in today's net savy world. The like is treated as a mark of appreciation and also a measure of popularity in cyber space. The like in social networks defines the extent of popularity and thus gives us a positive sentiment about recent trends in the market. Thus to harness this power of like the writer has proposed the methodology named as "Likeminer". Likeminer uses a heterogeneous computer network termed as 'LIKE network model', where each entity is treated as a node and the likes are given by attaching the edges in a directed graph. The higher the number of incoming edges to a node, the more popular will be the product. Three layers of extraction namely- object extraction, visual extraction and text

extraction are used in order to test the significance of likes. And a like mining algorithm is proposed to give us the most trending activities in cyberspace.

In their paper L et al [17] discuss about the ability to detect events over twitter. Seeing the trend of tweeting every single activity in a day has worked as a motivation behind this research paper. It suggests a web crawler interface which automatically monitors tweets flowing in the network and then analyzing these accessed tweets offline to get a sentiment score. The event chart is prepared and the list is checked with the data, matched entities gives us an event occurrence. This event occurrence may be geotagged hence giving s the entire area or vicinity where the event might have occurred.

The paper written by Pang et al. [18] deals in classifying documents not according to their topic but according to the sentiment value they contain. In other words it treats sentiment as a ranking measure for documents. The hypothesis taken in this paper is such that most of the documents are treated on basis of their keywords. And these keywords if give us a daring sentiment outcome, it can be used as a ranking mechanism. The author has used machine learning algorithms as SVM and bayes classifiers to get the sentiment of the documents and classify them according to their sentiment value. The unigrams and bigrams as calculated in order to define the results of the sentiment classification.

The survey paper written by Medhat et al. [19] tackles comprehensive overview of the last update in this field. Many recently proposed algorithms' enhancements and various SA applications are investigated and presented briefly in this survey. These articles are categorized according to their contributions in the various SA techniques. The related fields to SA (transfer learning, emotion detection, and building resources) that attracted researchers recently are discussed. The main target of this survey is to give nearly full image of SA techniques and the related fields with brief details. The most valuable outcome from this paper is a categorical classification of huge number of articles and setting up a trend in using sentiment analysis as a viable tool.

2.2 Keyword Extraction:

The paper written by Uzun [20] defines a mathematical formulae which can be used over a Bayesian classifier to extract a keyword from a phrase, or a set of documents. The mathematical formula suggested for implementation is tested on a varied set of documents to give an acceptable

keyword list. This paper very clearly defines the portion of text where the keyword may occurs and refuses to take frequency as the indicator of keyword determination. The mathematical formula based on conditional probability makes use of the TFxIDF score in order to identify a keyword in a document.

Turney [21] in his paper gives review of the various machine learning algorithms currently available which may be used as a method for extracting keywords from documents. The paper stars with defining a new term 'Keyphrase', as it seeks to find out a set of key words for processing in a document. The second part introduces the algorithms and how stemming is done in keywords in order not to take two variation of the same word as different keywords. The writer has for analysis taken five different document sets. A set of resultant outcome is also kept in order to exemplify the supervised learning methodology. Each document set is defined in a corpora and the rest of the functions are now applied to these corporas. Then various algorithms are applied and results are taken of the variation in the efficiency of these algorithms in finding out a keyword from the document.

Sarkar et al [22] gives a comparative study of three machine learning models, namely-naïve bayes, decision trees and artificial neural network. The document very elaborately explains how machine learning is used for keyword extraction, with the entire corpus formation and vector formation well defined. The second part deals with candidate keyphrase extraction and feature extraction models. These two models are used for the structure of extraction. Three type of datasets, namely-economics, law and medical documents are taken for testing the difference in these machine learning approaches. For benchmarking of the results the keywords defined in the documents are kept in a set and a dictionary is made. The dictionary is checked with the found keywords to give us the accuracy of different approaches.

Given a large text dataset composed of greater than 200k training sets and multiple classifications, various machine learning algorithms were used to train and predict tags and keywords.[23] The paper written by Zhang et al. [23] explores various techniques in pruning and managing a large, unwieldy dataset in order to produce a practical training point. Naïve Bayes and SVM are the algorithms focused on in this project, with various contours of the dataset tested to examine the practical effects of dataset manipulation.

2.3 Market Products:

Hootsuite [24] is an integrating tool, which integrates all your social media profiles to a single interface. It was introduced in 2008 by Ryan Holmes for the purpose of brand management of user profiles. Hootsuite gives a brilliant platform to summarize all the activities of a user on social media, and gives an analogy of the overall popular social feeds or an integrated opinion shared on multiple social networking platforms. Hootsuite started off as twitter dashboard named BrightKit. Hootsuite integrates all platforms of social media and gives us tools to analyze the overall activity in various pictorial and graph like interfaces. This is a very powerful tool which can be used by data miners to extract all sort of social media data coming out of individuals and make a consolidated sentiment analysis background.

WordNet [25] is considered as an ontology to give polarity for the words used. However, it was soon discovered that the relatedness among words is not symmetric. Words that may be considered related in disasters and disaster jargon (e.g., disaster and earthquake) show little WordNet relatedness.

SentiWordNet [26] is used to detect the sentiment value of Tweets. SentiWordNet uses lexcon based approach and is specifically used in sentiment analysis and opinion mining. Each word (or synset, in WordNet terminology) included in SentiWordNet has three characteristics: positivity, negativity and neutrality. SentiWordNet is based on first giving the lexicon sentiment values by a human being then using this as a seed to establish the values for other lexicons. Support vector machines were used to classify the lexicons linked to the seed. SentiWordNet is considered a general method to detect the sentiment in a text. It is found that some words have sentiment values even though they should be considered neutral e.g., here, a, sentiment, exist. A list of words that should not be considered in evaluating the sentiment value of the Tweet to avoid accumulating sentiment value for words that should be considered neutral.

2.4 Motivation

There are many papers and books regarding the evolution and use of social media and sentiment analysis in today's world. Some products have also been launched as mentioned above in order to make use of this social media realm. But there is no paper which sees to find the reason behind any negative public mood and negative sentiment regarding any topic. So suppose even if you get to know the public sentiment of any topic, but what is the significance of this analysis if we do not know the reason behind this public mood. This dissertation intends to find the public sentiment regarding a topic and also extracts the reason behind this public sentiment. The use of keyword extraction and feature extraction serves as the base for finding out the reason behind the negative sentiment. The TFxIDf score is taken to serve as a measure for the keyword identification.

CHAPTER-3
CHALLENGES

The world of social media is constantly evolving. In recent years social media has come up to become a giant data producer, and thus has contributed a lot to IT industry in terms of expediting the growth of data management and data manipulation. Thus this new field of social media mining sees a lot of innovation in data analysis. The data in these social networking sites is mostly and unstructured or semi structured. Thus the processing of data becomes a very tough and lengthy process. In order to make this easier data mining is used to extract information from these social websites. But the social front of these websites enables users to be in communities, and have different friend circles, thus making another tough task of monitoring and running through the entire network and its topologies. Some ardent challenges in social media mining are as follows:

3.1 Big Data Paradox:

The realm of social media is humongous. Thus the data generated as the outcome is huge. But from this huge data the amount of information is scarce. This altercation of huge data but scarce information is termed as 'Big data paradox'.[14] Simply speaking it is a tale like "Data Data everywhere, but no information to process". For solving this challenge of big data paradox we have used a search API, which uses a keyword as a base for searching for the data. Thus instead of data burst we get streamlined data regarding a specific keyword. We also use a language filter to streamline the data as, we use only English language for the data to be analyzed.

3.2 Obtaining Sufficient Samples:

Another challenge is how to decide that how many samples of data would be sufficient for obtaining a result.[14] We do not know the level of noise in the data, which makes it even more challenging to decide that are we even getting the opinion or not. Thus there is always a chance that the entire data that has been taken, might be meaningless and corrupt. Therefore deciding on the samples sufficient for knowledge extraction is a massive challenge. Thus for solving this challenge we take a constant number as a sample space. This sample space contains the number of data sets to be kept after streamlined through search API.

3.3 Noise Removal Fallacy:

Noise is the unrequired data that comes while analysis. Social media applications generate a lot of vague and noisy data.[14] Thus to remove the noise becomes mandatory. But how do we decide what is noise or not. For example: consider you want to analyze public mood of ‘abc corp’ a branch. Abc corp has multiple branches thus social media will capture everything. Now for you all the data of other branches excluding branch ‘a’ is noise. But in actual terms it is not noise. So this problem of noise recognition gives birth to the challenge of noise removal fallacy. Where noise is removed inappropriately or noise is not removed at all. Thus the keyword search scenario also solves the noise removal fallacy challenge as we get the filtered data for processing.

3.4 Deciding on keywords:

Another challenge is deciding on the keywords. As we seek to establish a reason for the negative public sentiment, we need to identify keywords. But in a set of data how to decide which of the words are keywords and which are not. Frequency alone cannot be a measure of keyword as prepositions are the most used words in a document, and they are seldom keywords. Therefore defining the measure for keyword definition is a major challenge. For solving this challenge we use keyword feature TFxIDF score of the words encountered along with the probability of the word being a keyword.

CHAPTER-4

BACKGROUND

4.1 Social Media:

Social media is a virtual space where all people can interact and monitor the progress of one another irrespective of the geographical bottlenecks faced by the people. The development of social media progressed hand in hand with the omnipotence of intenet. Social media provides us with the platform of getting a global perspective on our views. The core advantage of social media is its ability to engage people in conversations, no matter where they are. Due to this flag bearer of independent speech the geographical boundaries have diminished and we have witnessed a growth in each and every aspect of life. Thus social media has played a major role in developing businesses and foreign relations as well. Social media also indulges one and al in resource sharing and problem distribution making the solving of any problem easy.[14] The varied range of social media has contributed immensely in the growth of every aspect of our lives. For instance: Sites like facebook and Google + have made us interact with our long lost friends, cancelling the notion of distance among us. Technological Blogs and communities help us in solving our technical problems with good expertise available. Sites like twitter keeps us up to date about everything happening in our environment. Thus social media has settled deep into our daily lives within every vertical of our life.

4.2 Sentiment Analysis:

Sentiment in a general sense means mood or our emotions towards an entity. Thus sentiment analysis is basically to use a set of tools like natural language processing, computational linguistics and text mining to give us some knowledge about the mood regarding an event or object. Sentiment analysis is a subfield of data mining. Thus extracting knowledge from data is the core work of any sentiment analysis system, and this extracted information is actually the sentiment. With the growth of social media, the requirement to check the popularity charts and trend in social media, the use of sentiment analysis is to monitor public sentiment. [27] The best thing about analysis that it has to yield results. Thus sentiment analysis has to yield results in terms of sentiment classes. These classes can be positive, negative or neutral. Some superlative classes like very

negative or very positive may also be added to give a consolidated sentiment feedback. The analysis of sentiment also includes analyzing emoticons and the data it represent.

4.3 Data Mining:

Data mining is the process of extracting knowledge from data. It is also referred to as KDD or Knowledge discovery in databases. The core entity in data mining is data, therefore we have to first understand what data is. Data is a set of raw facts and figures. Data can be classified into three types, structured data, unstructured data and semi structured data. Structured data is the type of data which is related and is represented in tables in a database. Unstructured data is a bulk of data which cannot be put in a database table, because of the sheer bulk of representation units required. Semi structured data is data that is kept in tables but within tables it has multiple patterns and cannot be put into a classified relation. Extracting knowledge from all these type of data is a complex task indeed. It requires various activities like noise removal, consistency check, patterm matching to extract knowledge from data. [28] Data when processed is called information. Processing means, making the data relate to each other and make the sense out of an entity, thus relational data yields information. Information accumulated over long period of time generates knowledge. Thus data mining is a continuous process to extract useful information. For achieving this goal data mining use many subfields like machine learning, statistical learning and clustering to represent the data in a structured understandable mean.

4.4 Machine Learning:

In general terms, learning is a continuous process, it requires the understanding of our environment and the learning criterions. The seeds of machine learning are based on this mechanism only of how the human beings learn. We humans from our very child hood starts observing things and draw some metadata about every observation we make. Then we are trained to use these observations in our schools, and when our training is completed we are tested with examinations. Similarly machine learning is the procedure to make machine learn. In order to achieve this goal, the same procedure is followed, where the programmers feed a set of observation in term of code. Train the machine extensively and then test the machine for results.[29] But before testing the machine has to generate a pattern only then can the machine respond, because unlike human brain machine is not evolving on its own. Thus machine generates a pattern and then

extracts a mathematical formulae n order to create the answers on its own using the mathematical formulae. This entire process of machine generating result on its own is called as machine learning.

4.5 Keyword Extraction:

Keywords are considered to be some specific words, which stand out in a document or a group of words.[30] Getting to know which of the many words in a sentence is a keyword is a very tough task. Thus the process of identifying keywords in a group is termed as keyword extraction. Keywords alone can give the idea of the entire phrase about which it is talking about, thus to get to know the jist of a paragraph is easily possible if we identify the correct keywords. There are multiple techniques which can be used to identify the keywords. One such technique which we are using is feature extraction, where multiple features are taken into account to determine the keyword status.

4.6 Twitter:

Twitter is a social media website which enables its users to send short messages with length up to 140 characters. These messages are displayed publicly, and your friends and fellows can express their point of you by either supporting you by retweeting or commenting over your message. Tweets can only be posted by registered twitter users, whereas unregistered users can only read tweets. With the innovation in technology twitter now can provide several facilities like mobile recharges using tweets, such technology is known as hash tag money transfer. With the growth in the field of data mining and data analysis, twitter has become a good source of data for analysis, there are millions of users registered on twitter, and thus it serves as a good hunting ground for sentiment analysts. Twitter gives also a very good opportunity to developers interested to work on their data and allows them to create applications independently to be incorporated later to the twitter interface. Thus dev.twitter.com and apps.twitter.com is a much fancied site for developers and coders who want to use socially relevant data.

4.7 Twitter APIs:

For enabling and grooming youngsters to code, twitter has made its data available for scrutiny. This data can be taken as a .json fle. [31] The file needs to be accessed over a secure channel and thus twitter has created a couple of APIs to help users to work on their data and also

help twitter in developing. The APIs are developed by free communities and thus you can see and even alter the code. Some twitter APIs which are very useful to this project are as follows:

4.7.1 REST APIs:

The REST APIs are used to deliver a complete read and write access to programmers to work on twitter data. Any developer may use this API to read the twitter data from any profile and work on that data. It incorporates a security mechanism as well to give the developer a secure channel to work on this twitter data.[31] The OAuth utility gives this security mechanism by delivering a handshaking between the twitter developer and the website. Once the handshaking is complete the developer can access the tweets and Meta information on these tweets in order to get a much relied result. The responses of the OAuth are tracked as a json file, which gives the selected tweets and mechanism for working on twitter data.

4.7.2 The Search API:

REST API is a very good platform to work on the twitter data but when we need real time data from twitter REST is not the answer. The twitter developers have another utility named as search API to work on real time data. It is a part of REST API but has feature of timeline dependency incorporated. Thus the corroboration of tweets becomes trust worthy. This search API can be used just like a search tag in any website, where you can take a specific topic or keyword in terms of search and get the twitter data related to it. The search api searches for the relevant data but not the complete data so you might not have the complete result but you may have a good mean to work on this result. For a complete result one may use streaming API where the data keeps on streaming in as soon as the handshaking has been completed.

The search API uses a GET search option mechanism, where the request to be called is dened under GET and the response is recorded under Request. [32] Search tweets in general keeps tweets only for a week old tweets.

4.7.3 Authentication on all endpoints:

Search API and REST API both need a security mechanism to provide a secure channel for working on twitter data, thus authentication is required in each phrase of the communication. The multilayer authentication helps us to monitor the data sequences as well as provide a reliable

platform. [33] There are multiple checkpoints available to illustrate the unique secure connection, as each application is given a unique combination of keys in order to ensure the security. This access tokens and authorization codes generate a unique handshaking code every time the application has to link with twitter.

4.7.4 Application-user authentication

This authentication ensures the authenticity of the application communicating with twitter. [33] It is a part of OAuth v1.1. The application has unique codes of access and these unique codes then also later generate a unique sequence of numbers in order to ensure a user is authenticating the application. Thus the communication is now finalized after monitoring the type of data being saved and used.

4.7.5 Application-only authentication

In the above authentication procedure the application was authorized using the access tokens whereas the user was authorized on basis of the number sequence generated in the transition. But in application only authentication where only the application makes the API request, without actually considering the authentication of the user context. Here the basis of communication is not user, thus we can have information irrespective of the user as the information would be gathered only on basis of the application. Thus the multi-platform tweets may also be saved using this authorization mechanism.

4.7.5 Differences between Streaming and REST

The connection created using a Streaming API will make a permanent HTTP connection, which is open for all certain values. Thus the entire communication with the application differs because of the security mechanism changing. There is a basic architectural difference in the streaming and REST API.

An application which is created using Streaming API will not let the user autonomously establish a connection on a request basis. But the code is run to generate a continuous app request to work with the twitter data. The connection call the code and the session will run till the time the GET is receiving the incoming messages. Figure 4.1[33] shows The entire process of setting up a streaming API:

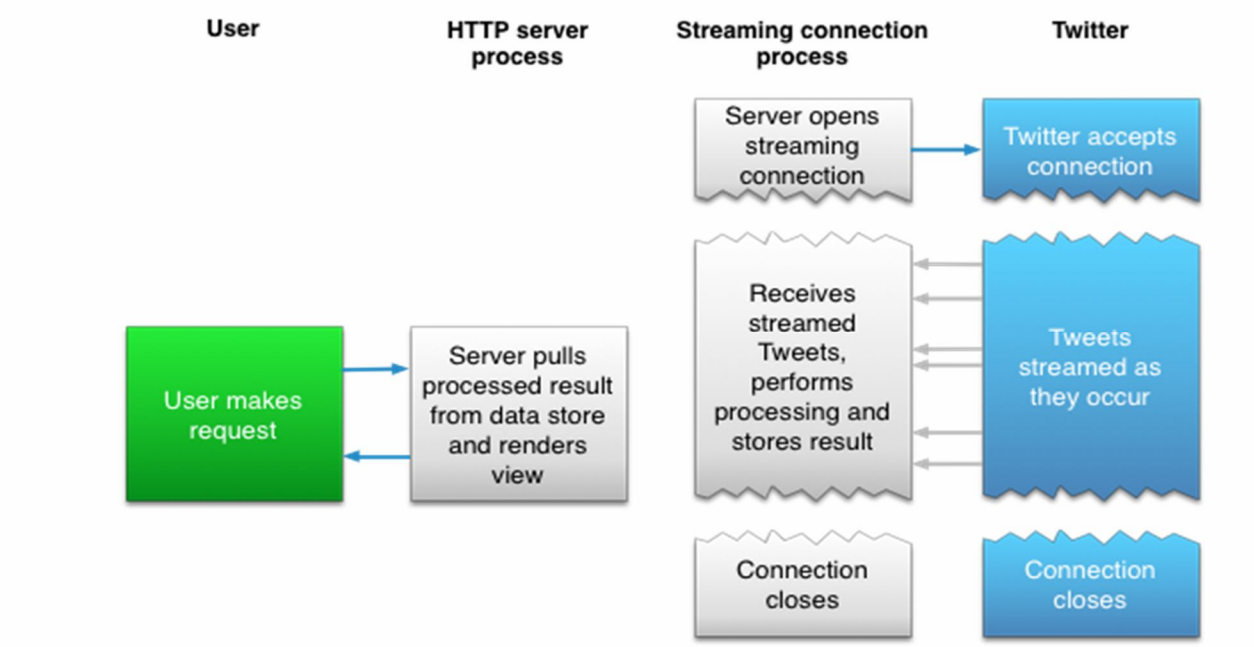

Fig4.1: Architecture for streaming API

The streaming process gets the input Tweets and performs any parsing, filtering, and/or aggregation needed before storing the result to a data store. The HTTP handling process queries the data store for results in response to user requests. While this model is more complex than the first example, the benefits from having a realtime stream of Tweet data make the integration worthwhile for many types of apps. Fig 4.2 [33] explains the complete architecture of REST API.

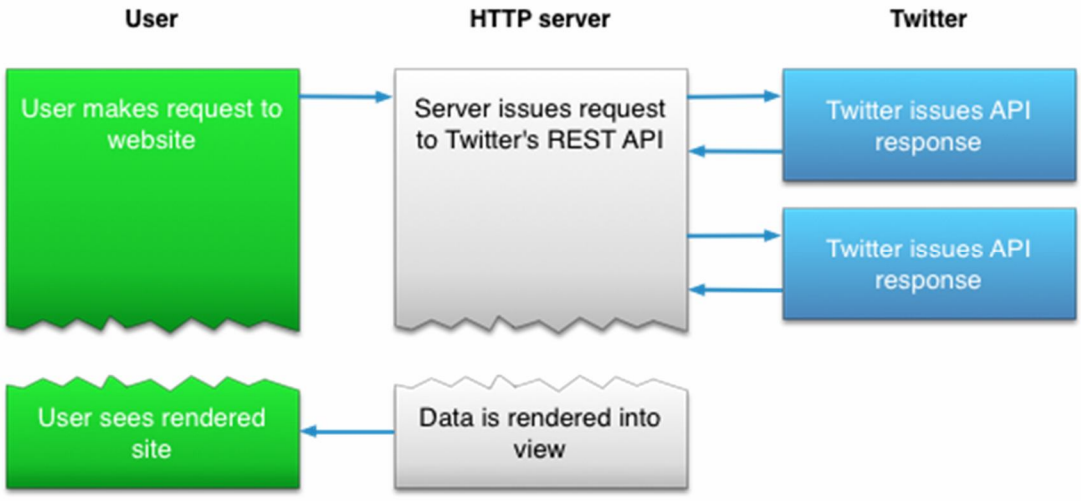

Fig4.2: Architecture for REST API

CHAPTER - 5

REQUIREMENTS

5.1 Hardware Requirements:

- 6GB of RAM, including swap space.
- 2.2GHz processor. (Additional processing power may be required for multiple concurrent styling renderings)
- 1 GB software disk usage.
- 64-bit hardware recommended.

5.2 Software Requirements:

- Operating Systems: Ubuntu 12.04 or Windows8
- RStudio 0.98.507

5.3 Packages Required:

5.3.1 **TwitteR:** Provides an interface to the Twitter web API.[34] This function will issue a search of Twitter based on a supplied search string.

- **searchString:** Search query to issue to twitter. Use "+" to separate query terms.
- **n:** The maximum number of tweets to return
- **lang:** If not NULL, restricts tweets to the given language, given by an ISO 639-1 code
- **Geocode**: If not NULL, returns tweets by users located within a given radius of the given latitude/longitude.
- **maxID:** If not NULL, returns tweets with IDs smaller (ie older) than the specified ID. These commands will return any authorized tweets which match the search criteria. Note that there are pagination restrictions as well as other limits on what can be searched, so it is always possible to not retrieve as many tweets as was requested with the n argument. Authorized tweets are public tweets as well as those

protected tweets that are available to the user after authenticating via registerTwitterOAuth.

5.3.2 plyr: The plyr package is a set of clean and consistent tools that implement the split apply-combine pattern in R.[35] plyr is a set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each pieces and then put all the pieces back together. For example, you might want to fit a model to each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arrays to simpler summary statistics. The development of plyr has been generously supported by BD (Becton Dickinson).

5.3.2.1 Count: Speed-wise count is competitive with table for single variables, but it really comes into its own when summarizing multiple dimensions because it only counts combinations that actually occur in the data. Compared to table as.data.frame, count also preserves the type of the identifier variables, instead of converting them to characters/factors.

5.3.2.2 laply: Split list, apply function, and return results in an array. For each element of a list, apply function then combine results into an array.

5.3.3 stringr: stringr is a set of simple wrappers that make R's string functions more consistent, simpler and easier to use. It does this by ensuring that: function and argument names (and positions) are consistent, **str_split** Split up a string into a variable number of pieces. Vectorised over string. Pattern should be a single pattern, i.e. a character vector of length one.

5.3.4 RCurl: The package allows one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server. This provides a great deal of control over the HTTP/FTP/... connection and the form of the request while providing a higher-level interface than is available just using R socket connections. Additionally, the underlying implementation is

robust and extensive, supporting FTP/FTPS/TFTP (uploads and downloads), SSL/HTTPS, telnet, dict, ldap, and also supports cookies, redirects, authentication, etc.

5.3.5 ROAuth: Class OAuth wraps and handles OAuth handshakes and signatures for the user within R. Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice. Version 0.9.6. The OAuth class is currently implemented as a reference class. An instance of a generator for this class is provided as a convenience to the user as it is configured to handle most standard to access this generator, use the object OAuthFactory. See the examples section below for an example of how to instantiate an object of class OAuth. In almost all cases, saving an OAuth object after handshake and loading it into future sessions will allow it to remain authorized without needing any manual intervention that might have been performed initially, such as the PIN step with Twitter authentication. Use the function save to save the credential object to a file and then load in another R session to bring it back in - there should be no reason to undergo another handshake by doing this. The needsVerifier argument is optional and defaults to TRUE. In almost all cases, the default should be used, the option is primarily provided to enable the examples as the keys provided by the examples are already signed. If you feel that you're in a situation where this should be set to FALSE, it's best to double check this. The signMethod to the handshake method tells the system which OAuth signature hash to use, one of HMAC for HMAC-SHA1 (default), RSA for RSA-SHA1 (not implemented), or text for plaintext. The customHeader argument to OAuthRequest can be used to pass additional HTTP header commands to the underlying request. The curl arguments can be used to provide a custom curl header, defaulting to a generic getCurlHandle call.

5.3.6 Rjson: JSON (JavaScript Object Notation) is a lightweight data-interchange format. This package converts JSON objects into R objects and vice-versa.

5.3.7 Tm: A framework for text mining applications within R. It contains methods for data import, corpus handling, preprocessing, metadata management, and creation of term-document matrices. The main structure for managing documents in tm is a so-called

Corpus, representing a collection of text documents. A corpus is an abstract concept, and there can exist several implementations in parallel. The default implementation is the so-called VCorpus (short for Volatile Corpus) which realizes a semantics as known from most R objects: corpora are R objects held fully in memory.

5.3.8 Snowballc: An R interface to the C libstemmer library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary.

5.4 Representation:

A data frame is a storage class representation of variables. It contains a matrix like structure with a number of rows with unique row names, and a number of columns. The rows in general shows the entities whereas the columns keep the values of these entities. The duplication of column names are allowed in a dataframe. A dataframe stores tables of data. It contains vectors of equal length. The column names are kept into a header. The data entries in a dataframe are kept in data cells where the row and column coordinates gives us the data value stored in a data cell. A data frame is just like a table in relational database.

CHAPTER-6
METHODOLOGY

For achieving our objective, the methodology is divided into three parts, sentiment analysis, and preprocessing and keyword extraction. But the very first agenda is to communicate with the twitter app. Fig. 6.1 explains the communication between host and twitter app.

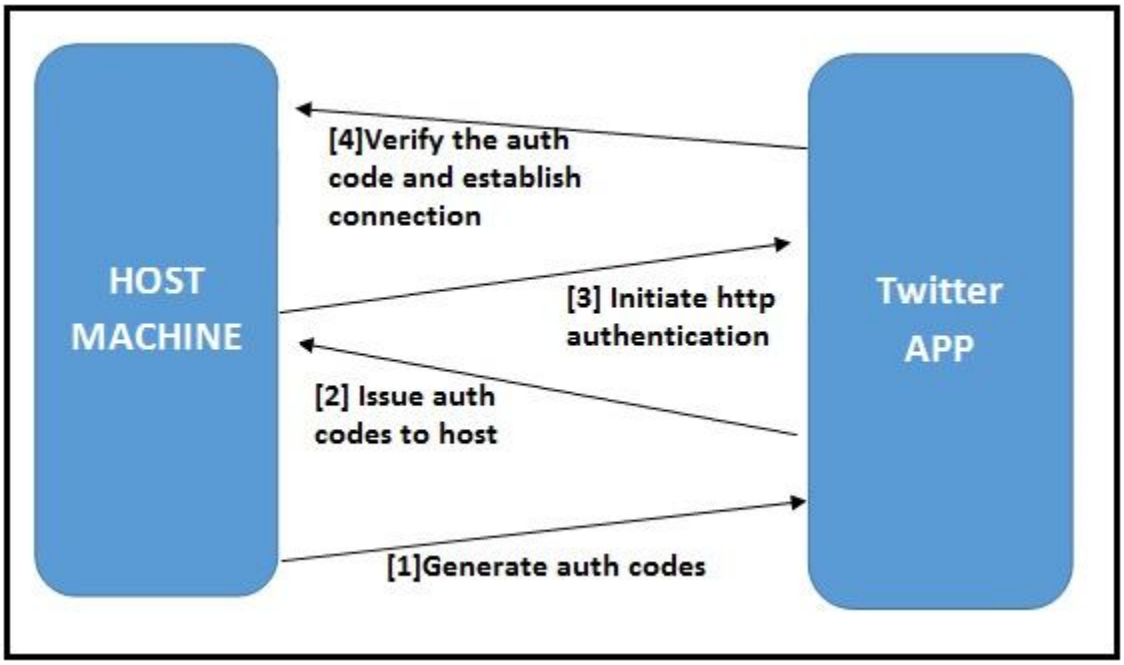

Fig. 6.1: Handshaking Process between Twitter App and host

The host will log into twitter app and create an application. With the creation of application the host will send a request for auth codes to twitter API. In response Twitter API will deliver auth codes to the host. The host will store these auth codes and initialize a session request on basis of these auth codes. The session request will then be sent to twitter api which will, require the app and auth code validation. Once the validation is complete, the session will be established between app and host. Once the communication is established, host can download tweets from the twitter api with any specific requirement and then start sentiment analysis on the downloaded tweets.

6.1 Sentiment Analysis: For analyzing the sentiment of social media data we need to follow below steps:

1. Try and download the social media information in a system readable format (.xls file or a data frame).
2. The downloaded information needs to be analyzed and a proper sentiment is associated with the information.
3. Associate the geological information of the social media data with the map. The entire process is classified into three sections.

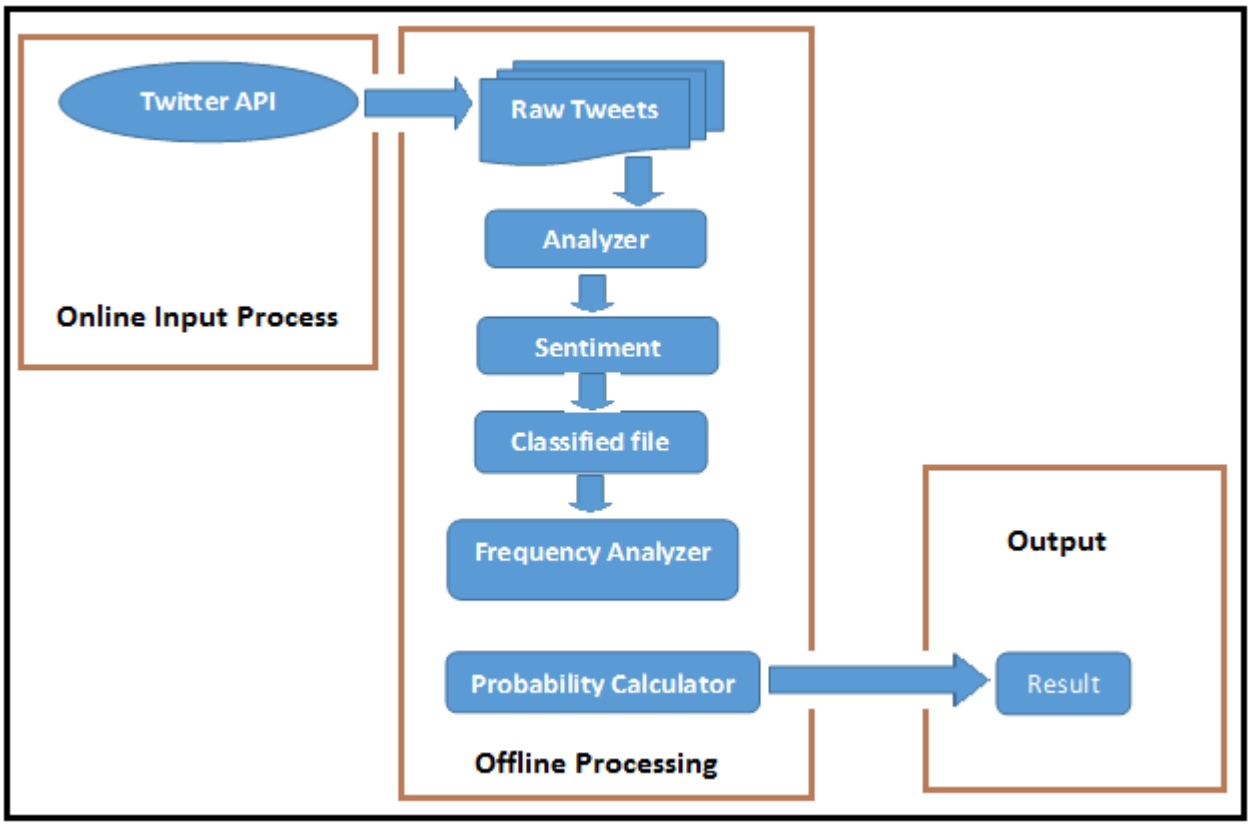

Fig6.2: Reference Architecture and Overview

The online process deals with the entire scenario of creating twitter APIs and handling the twitter feed, until a consolidated data frame has been formed. The offline process deals with generating the class labels for the tweets to be assigned and extracting meta information from the tweets. The analyzer then overtakes the process and gives us the sentiment files, which leads to our resultant data.

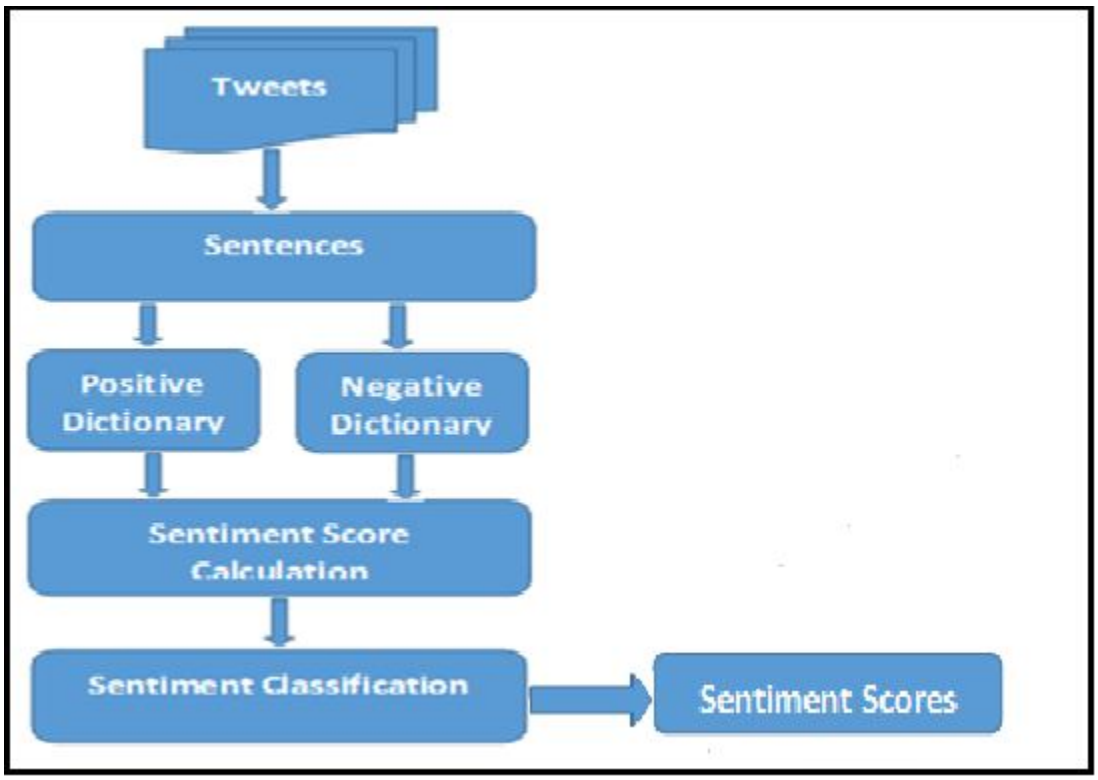

Fig6.3: Analyzer Architecture

Figure 6.2 shows the overall architecture of the analyzer system. We start by loading the tweets, a positive sentiment dictionary and a negative sentiment dictionary. After loading the lists we match and search for the sentiment value of the words in Tweets as a result. Afterwards we express the sentiment in a mesh and classify them into our three sentiment classes, positive, negative or neutral. We start by preprocessing the Tweets. We remove the punctuation marks and the stop words. We use the list based on google, the open source search engine. Then the tweets are tokenized. The next layer is responsible for matching the tokens of the tweets against the positive sentiment dictionary and negative sentiment dictionary. The sentiment score for the Tweet is accumulated. Finally, a value for the sentiment of the Tweet is calculated depending on whether it is positive, negative or neutral. Now the negative file will be taken as the testing data set. The reason for this negativity is further termed as key.

6.2 Preprocessing the Documents:

Extracting a set of keywords from a document is a very technical job. Each word in a document does not has the same probability of becoming a keyword. Therefore a lot of document preprocessing is required for the document to me working as a data source for keyword extraction.

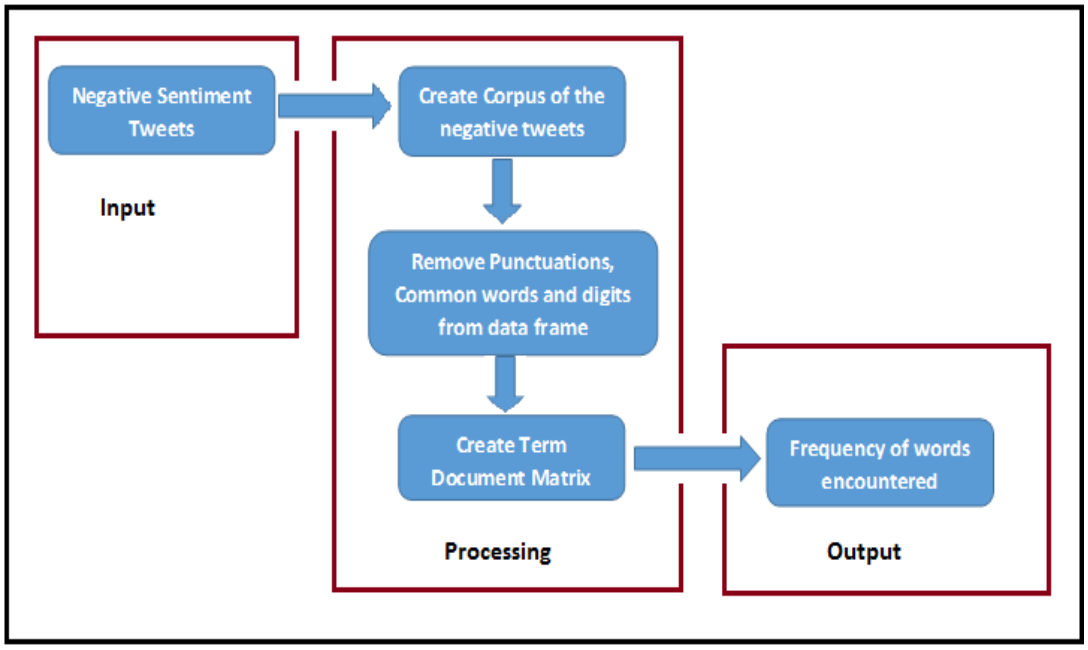

Fig 6.4: Preprocessing For Keyword Extraction

The first process is to delimit the word wherever a space, a tab, an enter or a dot has occurred. The frequency of words in a document may also be a criterion for stating it as keyword. But frequency of prepositions is also the highest in any document but they do not serve as a capacity of keyword. Thus we can remove the prepositions from the document. Thus now our document seems to be ready for further processing and extracting keywords.

6.3 Keyword Extraction:

There might be thousands of words in a document and searching for keywords among these thousands of words. Thus we need to add some utility to understand and extract keywords from any document. One such utility is feature extraction. Each keyword may be weighed in terms of various features. These features might be the frequency of word in a document, the placing of the word in the document and the repetition of the word in a paragraph and so on. These many features

can enable us to solve a very complex task of extracting the important words in the document. One such very important feature is the TFxIDF score, Which stands for term frequency x inverse document frequency score.

$$\text{TFxIDF}(x) = P(x) \times [\, -\log P(y) \,].$$

$$P(x) = \frac{f(x)}{\max f(x)}$$

$$P(y) = \frac{f(x)}{D}$$

Where,

f (x) is the number of times a word has occurred in the document

max f(x) is the maximum frequency of a word

D is the total number of words in the document.

One of the major problems in using this feature is that if a word has been repeated multiple times the second portion will always tend to zero. Therefore the training should be taken such that it contains a lot of words to maintain a balance in the occurrence in the document.

Thus based on the multiple features explained a slight variation to Bayes theorem can be a very good idea to give us the keywods we are looking for:

$$P(\text{Keyword\_score}) = \text{TFxIDF}(x) * P(\text{Keyword})$$

The most probable keyword would be having the highest P(Keyword score). A threshold could be defined in order to get a better set of keywords.

CHAPTER-7

IMPLEMENTATION

We have used Search API of twitter with extensive use of R to analyze the tweets. The basic steps are as follows:

- Creating an app in twitter to search for the tweets
- Connecting the analysis tool R with twitter
- Download the tweets to a csv file
- Read the csv file to match with dictionary and give the sentiment.
- Make the summary of the sentiment scores to give the overall public sentiment of the topic

7.1 Setting up Twitter API

1. Login to apps.twitter.com and create a twitter App
2. Generate API Secret Key
3. Generate API Key
4. Provide read and write permission to the API
5. Generate Consumer Key
6. Generate Consumer Secret Key
7. Generate Auth Tokens
8. Generate Access Token

The sequence of these steps, creates a twitter search API which is later used to communicate with the mining engine. The OAuth tokens used for a secure communication with the mining engine are established.

7.2 Connecting With the Twitter API

1. Store access URL, reqURL, authURL, consumer key and consumer secret to a variable.
2. Initiate a new handshake using the defined variables.
3. Store cacert.pem certificate for secure handshake
4. Verify the session with the app security code.
5. Register the session

6. Search for the tweets and save the tweets to a dataframe.

Once the secure codes are created in the twitter app, the mining engine starts to communicate with the twitter app, by using these secure codes. It then checks for the necessary cacert certificate to complete the handshake process and establishes a session. This session now holds, till the mining engine or the app is not closed.

ALGORITHM 7.1 – Calculating the sentiment score

```
Calculate score{
  loop till EOD (a, positive word dictionary, negative word dictionary, .progress='none')
      {
       values := Loop for all the sentences encountered (a, function(b, positive word
                  dictionary, negative word dictionary)
       {
              remove punctuations from sentence 'b'
              remove controle structures from sentence 'b'
              remove digits from sentence 'b'
              tolower (b)
              list := split strig b on the occurrence of space and store the element
               w := unlist (list)
               positive match := match(w, positive word dictionary)
              negative match := match(w, negative word dictionary)
              value := sum(positive match) - sum(negative match)
               return(value)
        }, positive word dictionary, negative word dictionary, .progress=.progress)
       return(values)
       }
```

Fig 7.1 Sentiment Analyzer Algorithm

Once the communication is validated, the analyzing algorithm takes over the process, and searches for a specific keyword related tweets, with specified language and sample space constraints. These tweets are then saved as a data frame and the algorithm 7.1 is run, which

associates a sentiment score with each tweet. And then summarizes the entire tweet score to give us the overall public mood regarding that topic. The negative set of tweets is then carried out to another data frame, which works as an input for the frequency analyzer script.

ALGORITHM 7.2 – Remove Operation

```
Remove {('entity', from which sentence to remove (a), ignore.case := FALSE, perl :=
        FALSE, fixed := FALSE, useBytes := FALSE)
{
  if (!is.character(a))
  a <- as.character(a)
}
```

Fig 7.2: Algorithm for remove operation

The remove operation algorithm takes a set of values and removes the values which are not characters. As the special symbols and digits establish no sentiment value, thus all such elements like punctuation, digits and control structures are removed from the data frame.

ALGORITHM 7.3 – Matching Operation

```
Match {(x, table, no_match := NA_integer_, incomparables := NULL)
        loop(x!=EOD)
        if (x[], table)
        {
            Returm(match)
        }
```

Fig 7.3: Algorithm for matching the elements

The matching algorithm 7.3 takes two lists and compares the two lists. If the elements of the list matches the reference list it returns a logical true, and if not returns false. The match operation does not allow assignment of values in the lists.

ALGORITHM 7.4 – Unlisting Operation

```
Unlist{ (x, recursive = TRUE, use.names = TRUE)
        {
        if (loop(islistfactor(x, recursive)))
        {
        level := unique(unlist(x, levels), recursive, FALSE))
        name := if (names=match list)
                names(unlist(x, recursive, matchlist))
        result := (unlist(x, as.character), recursive, FALSE))
         result := match(result, level)
        structure (result, levels = level, names = name, class = "factor")
        }
       else (unlist(x, recursive, names))
}
```

Fig 7.4: Algorithm for unlisting the elements

The unlisting algorithm 7.4 takes words out from a string of words. It takes the list of words and factors the list recursively until the end of document is achieved. The unlist works over the matching mechanism where the space works as the delimiter, and when a space is encountered the word is popped and compared with a list of words, if the match is found then the word is popped out else the loop continues for the match. It takes as input and entire string of words and return a dataframe with words.

After the words have been unlisted from the document, the frequency of the word, meaning how many times a word has occurred in the entire document is to be calculated.

7.3 Calculating the word frequencies

1. **Input:** A data frame with negative sentiment tweets.
2. Create a corpus of the data frame.
3. Remove all punctuations from the corpus.
4. Remove all digits from the corpus.
5. Remove unnecessary words from the corpus.
6. Change the corpus to lower case.
7. Create term document matrix from the corpus.
8. List all non-zero elements of the matrix.

For calculating the frequency of words, the first step is to create a corpus. A corpus is a set of words which is taken as a word listing. The corpus can be created by taking the document as a vector string with each vector giving the word. Then remove all unnecessary information like punctuations and digits from the corpus and convert the corpus to a lower case for uniformity. Once the corpus is defined, create a term document matrix. This term document matrix is a square matrix with the rows and column holding the word and the element in the matrix defining the count of word being repeated. All non-zero elements are popped out of the matrix and stored in a data frame with their frequencies lying in the deceasing or ascending order.

7.4 Calculating the Keyword score

1. Input: Word frequency list in decreasing order.
2. Generate the word probability.
3. Generate the Word document probability.
4. Calculate the inverse log values for word document probability.
5. Generate TFxIDF score for each word.
6. Match the frequency list with the keyword probability list.
7. Create the keyword probability score.

Once the frequency document has been finalized the feature extraction of the document starts.

CHAPTER-8

RESULTS

We collected 1000 tweets of the keyword “Disaster” and after running the sentiment analyzer we found 640 negative, 87 positive and 273 neutral tweets.

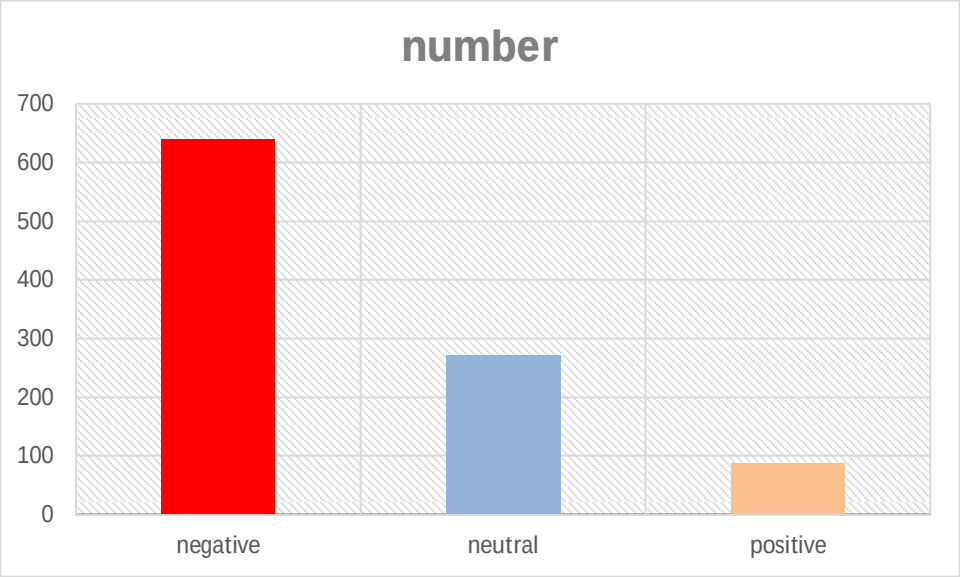

Fig8.1: The sentiment chart

These Negative tweets were then taken as an input file and it was found that there are total 10096 words in this file. The word length of each sentence was also calculated. On basis of the mathematical model a list of probable keywords was extracted as:

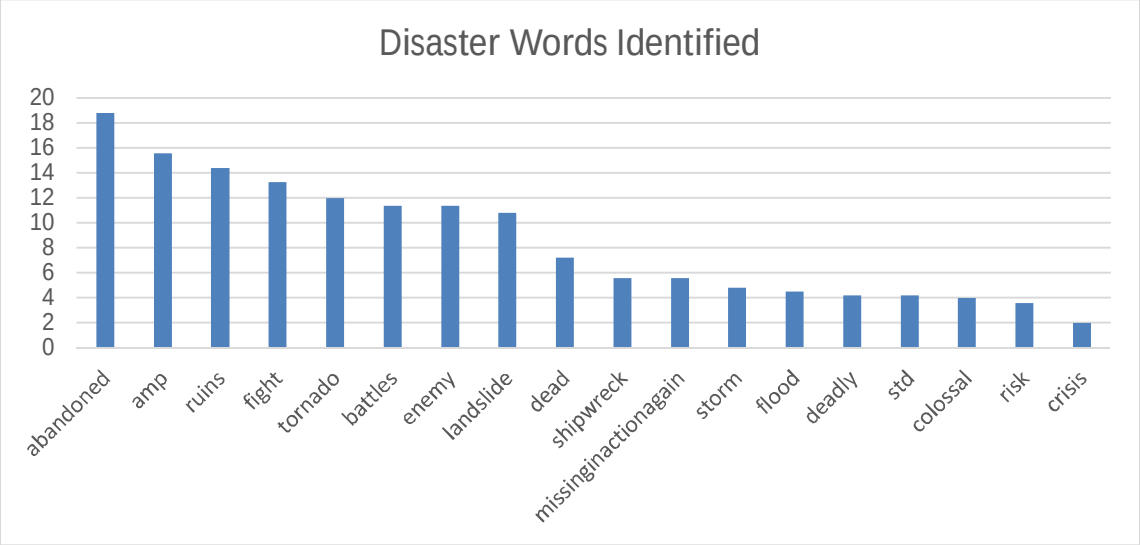

Fig8.2: The probable keywords

Similarly the tweet feed was checked for more results. I downloaded 1000 tweets with the keyword “Yemen” and 2000 tweets with the keyword “Nepal”. The summary of the tweet feed is evident in Fig8.3 and Fig 8.4.

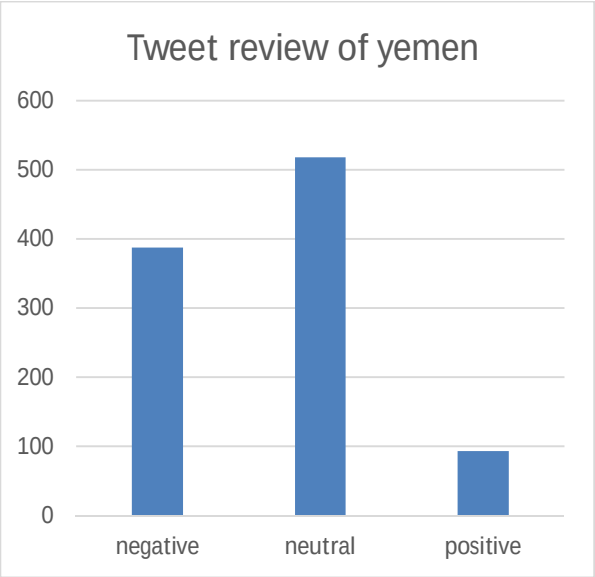

Fig 8.3: Sentiment analysis of yemen

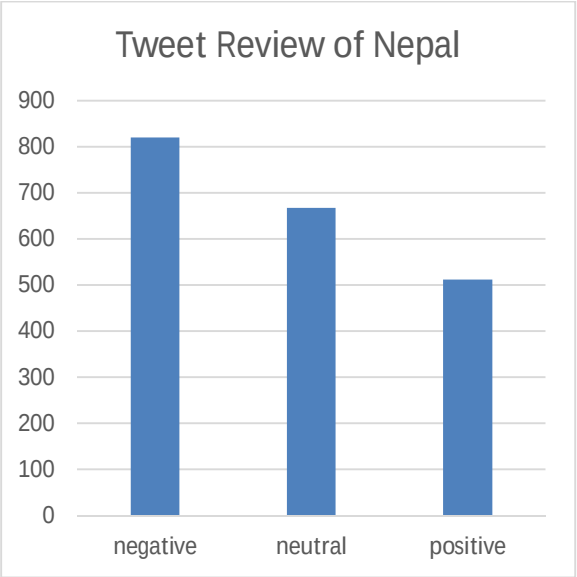

Fig 8.4: Sentiment Analysis of Nepal

The figures above show that the public sentiment of Yemen and Nepal is very negative. Thus after the analysis of the sentiment we found out the probable reasons for this negative feed. Fig 8.5 and Fig 8.6 show the reason of negative sentiment of Yemen and Nepal respectively.

count	word	tf_score	word_prob	log_values	TFxIDF_scc	Keyword_F	Keyword_Score
121	strikes	0.333333	0.019457	3.939574	1.313191	0.5	0.656596
86	deadly	0.236915	0.013829	4.281017	1.014235	0.7	0.709965
86	rebel	0.236915	0.013829	4.281017	1.014235	0.7	0.709965
34	war	0.093664	0.005467	5.209004	0.487896	0.7	0.341527
32	dead	0.088154	0.005146	5.269628	0.46454	0.9	0.418086
21	conflict	0.057851	0.003377	5.690842	0.329222	0.6	0.197533
8	killed	0.022039	0.001286	6.655923	0.146687	0.9	0.132018
7	crisis	0.019284	0.001126	6.789454	0.130926	0.8	0.104741
7	kill	0.019284	0.001126	6.789454	0.130926	0.4	0.05237
6	bombing	0.016529	0.000965	6.943605	0.11477	0.8	0.091816
6	houthis	0.016529	0.000965	6.943605	0.11477	0.8	0.091816
4	militants	0.011019	0.000643	7.34907	0.080981	0.8	0.064785
3	intolerable	0.008264	0.000482	7.636752	0.063114	0.4	0.025245
3	violence	0.008264	0.000482	7.636752	0.063114	0.7	0.04418
2	chaos	0.00551	0.000322	8.042217	0.04431	0.3	0.013293
2	crash	0.00551	0.000322	8.042217	0.04431	0.8	0.035448
1	fight	0.002755	0.000161	8.735364	0.024064	0.7	0.016845
1	unrest	0.002755	0.000161	8.735364	0.024064	0.6	0.014439

Fig 8.4: Reason of negative feed in Yemen

count	word	tf_score	word_prob	log_values	TFxIDF_scc	Keyword_F	Keyword_Score
471	earthquak	0.672857	0.03527	3.344713	2.250514	0.9	2.025463
123	missing	0.175714	0.009211	4.687387	0.823641	0.6	0.494185
52	killed	0.074286	0.003894	5.548328	0.412161	0.9	0.370945
14	lost	0.02	0.001048	6.860514	0.13721	0.6	0.082326
11	dead	0.015714	0.000824	7.101676	0.111598	0.9	0.100438
9	deadly	0.012857	0.000674	7.302347	0.093887	0.7	0.065721
8	crisis	0.011429	0.000599	7.42013	0.084801	0.8	0.067841
7	tragedy	0.01	0.000524	7.553661	0.075537	0.7	0.052876
6	hospital	0.008571	0.000449	7.707812	0.066067	0.5	0.033033
3	strikes	0.004286	0.000225	8.400959	0.036004	0.5	0.018002
3	war	0.004286	0.000225	8.400959	0.036004	0.7	0.025203
1	chaos	0.001429	7.49E-05	9.499571	0.013571	0.3	0.004071
1	risk	0.001429	7.49E-05	9.499571	0.013571	0.6	0.008142

Fig 8.6: Reason for negative feed in Nepal

CHAPTER-9
CONCLUSION

Social media has grown rapidly in the last decade and it continues to grow by leaps and bounds. Thus our system realizes the importance of this tool in our daily lives and try and make a system which can help us in deciding a public sentiment. We have seen the evolution and various technologies associated with social media and how our framework helps in accessing the social web with an analytical point of view. The results show that we have been partially successful in calibrating the social feed. Now these keywords that we have achieved gives us an idea of the negative public sentiment. We can observe from our results that out of 1000 tweets downloaded about Yemen nearly 400 are negative and 100 are positive, thus we conclude that there is a negative public mood about Yemen. After the analysis we can see the result that this negative mood is a result of a deadly rebellion, which has created a war like scenario in Yemen. Similarly out of the 2000 tweets analyzed for Nepal, we found that there are nearly 800 negative tweets and 500 positive tweets, thus the public mood regarding Nepal is also negative. The reason for this negative sentiment about Nepal is found to be a deadly earthquake which has created a state of crisis. A differential analysis of Nepal also concluded that the relief work in Nepal is on the rise, and thus the positive sense is continuously rising, due to worldwide efforts. But there are some slight limitations also associated with the work done, as we are using a REST API thus we have a limit to the number of tweets we can access and even the dimensionality of the tweets is reduces. But still it remains a very good source of information extraction. The analytics in social media has just become and are a prospect of future. Thus this project also incorporates various future possibilities in term of a app, which keeps you updated about the traffic scenario, or the rush to hill stations while on holiday, or even a very grave topic of disaster evaluation. We can also add a GIS system over this system to find out the exact nature and feed output and relevance. Thus this sentiment analysis system might prove to be a stepping stone for various advanced applications to run. This is a very basic framework, which we have successfully tested, but there are a lot of possibilities over this framework.

9.1 Future Possible Improvements:

Sentiment analysis is a very new domain and my attempt is just a beginning to a much bigger picture. My work has just been able to design a basic framework for a general purpose sentiment analyzer and finds the reason for this negative sentiment. But there are multiple improvements possible to this framework:

1. The matching dictionary and word probability is manually defined, but an autonomous assignment of probability to occurring words would highly improve the accuracy of the system.
2. The framework can be highly effective for disaster evaluation, where a specific location can be tracked for tweets, and the matching dictionary contain the various places of the location, now if any disaster has taken place in that area, the info can be represented in a GIS system and give us a real time data of the effect and extent of disaster in that area.
3. The framework can also be used for stock investment guide. If an expert system be designed over this framework which gives us a guide to which stocks to invest into on basis of their public perception and the reason for their public perception.
4. One important open area is to device a communication protocol which gives us an on call access to tweets. The twitter APIs currently available either close on with the session expiration and the reconnecting is again to be done from scratch or the http port is made open and thus making the app viable for attacks, thus a requirement for login based session may give us a more effective communication.
5. Deciding on the threshold for keyword validation is also an open issue. We have used a manual approach to decide the threshold on base of the keywords, but a system is possible where the threshold be dynamically changed according to the frequency set and the topic and the amount of neutrality it has generated.

ANNEXURE-A

(Application Snapshots)

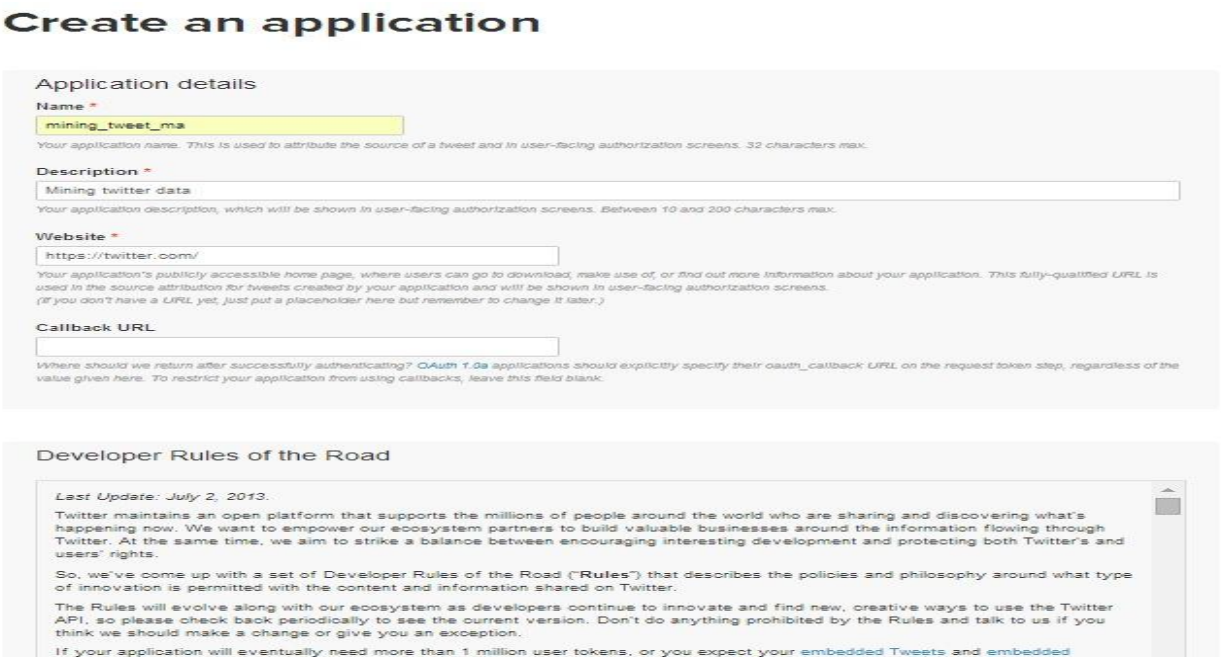

Fig A1: Twitter application creation

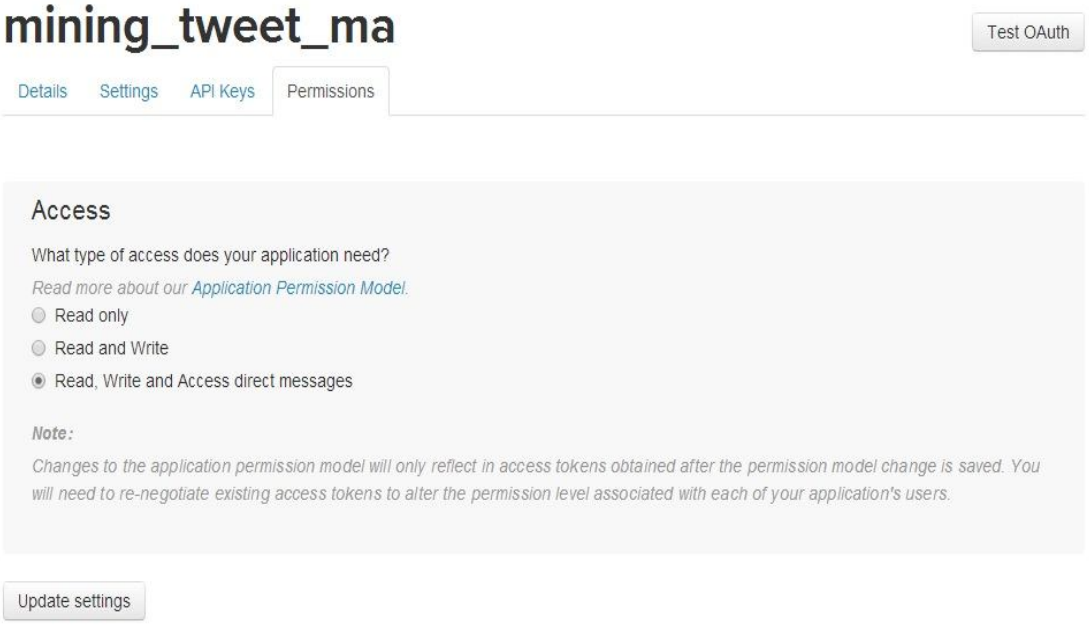

Fig A2 : Changing Twitter application permission

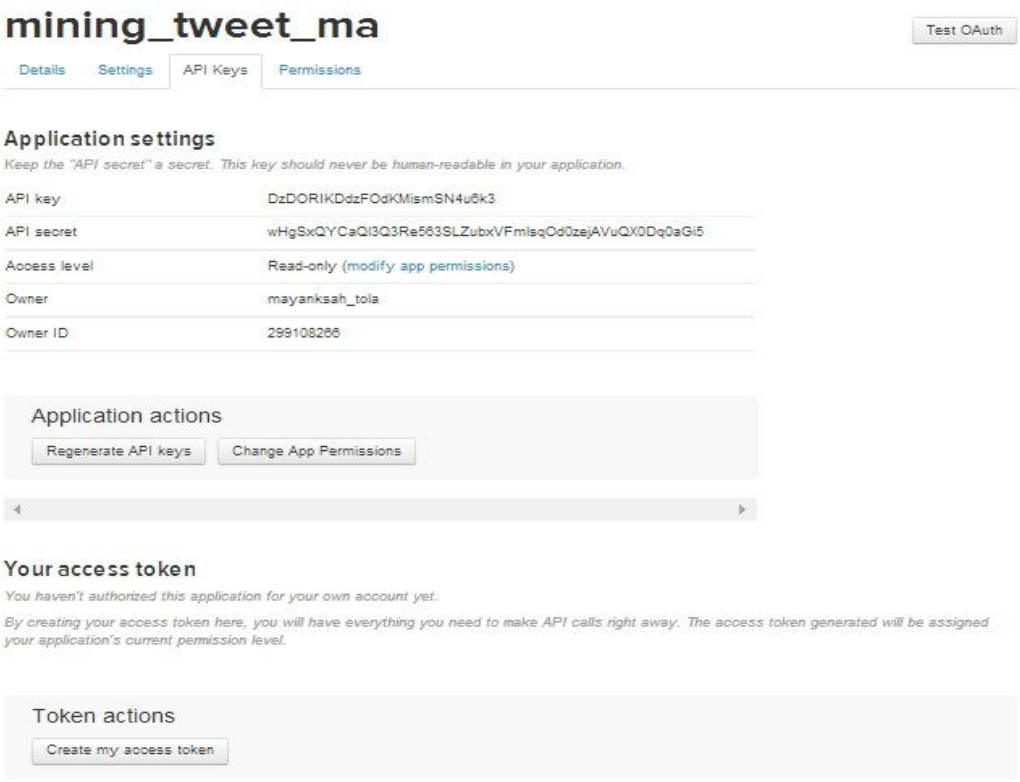

Fig A3 : Creating Twitter application access tokens

REFERENCES

[1]. Abdullah and Yunus. "Yunus Abdullah-Use of Social Media by Businesses: A New opportunity For Consulting Services by Accounting Firms." (2013)..

[2]. Goble and Gordon. "The history of social networking." Digital Trends 6 (2012).

[3]. Kononenko, Igor and Matjaž Kukar. Machine learning and data mining. Elsevier, 2007.

[4]. Zafarani, Reza, Mohammad Ali Abbasi and Huan Liu. Social media mining: an introduction. Cambridge University Press, 2014.

[5]. Ingale, Vijayshri R. and Rajesh Nandkumar Phursule. "Sentiment Analysis by Visual Inspection of User Data from Social Sites-A Review on Opinion Mining."

[6]. Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau. "Sentiment analysis of twitter data." In *Proceedings of the Workshop on Languages in Social Media*, pp. 30-38. Association for Computational Linguistics, 2011.

[7]. Hogenboom, Alexander, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong and Uzay Kaymak. "Exploiting emoticons in sentiment analysis." In*Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pp. 703-710. ACM, 2013.

[8]. Choi, Yoonjung, Youngho Kim and Sung-Hyon Myaeng. "Domain-specific sentiment analysis using contextual feature generation." In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pp. 37-44. ACM, 2009.

[9]. Maks, Isa and Piek Vossen. "A verb lexicon model for deep sentiment analysis and opinion mining applications." In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 10-18. Association for Computational Linguistics, 2011.

[10]. O'Hare, Neil, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin and Alan F. Smeaton. "Topic-dependent sentiment analysis of financial blogs." In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pp. 9-16. ACM, 2009.

[11]. He, Yulan. "Incorporating sentiment prior knowledge for weakly supervised sentiment analysis." *ACM Transactions on Asian Language Information Processing (TALIP)* 11, no. 2 (2012): 4.

[12]. Russell, Matthew A. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. "O'Reilly Media, Inc.", 2013.

[13]. Thiel, Killian, Tobias Kötter, Michael Berthold, Rosaria Silipo and Phil Winters. "Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining." *KNIME.com Report* (2012).

[14]. Zafarani, Reza, Mohammad Ali Abbasi and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[15]. Liu, Bing. "Sentiment analysis: A multi-faceted problem." *IEEE Intelligent Systems* 25, no. 3 (2010): 76-80.

[16]. Jin, Xin, Chi Wang, Jiebo Luo, Xiao Yu and Jiawei Han. "LikeMiner: a system for mining the power of'like'in social media networks." In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 753-756. ACM, 2011.

[17]. Li, Rui, Kin Hou Lei, Ravi Khadiwala and KC-C. Chang. "Tedas: A twitter-based event detection and analysis system." In *Data engineering (icde), 2012 ieee 28th international conference on*, pp. 1273-1276. IEEE, 2012.

[18]. Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79-86. Association for Computational Linguistics, 2002.

[19]. Medhat, Walaa, Ahmed Hassan and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* 5, no. 4 (2014): 1093-1113.

[20]. Uzun, Yasin. "Keyword Extraction Using Naïve Bayes." In *Bilkent University, Department of Computer Science, Turkey www. cs. bilkent. edu. tr/~ guvenir/courses/CS550/Workshop/Yasin_Uzun.pdf*. 2005.

[21]. Turney, Peter D. "Learning algorithms for keyphrase extraction." *Information Retrieval* 2, no. 4 (2000): 303-336.

[22]. Sarkar, Kamal, Mita Nasipuri and Suranjan Ghose. "Machine Learning Based Keyphrase Extraction: Comparing Decision Trees, Naïve Bayes, and Artificial Neural Networks." *JIPS* 8, no. 4 (2012): 693-712.

[23]. Zhang, Kuo, Hui Xu, Jie Tang and Juanzi Li. "Keyword extraction using support vector machine." In *Advances in Web-Age Information Management*, pp. 85-96. Springer Berlin Heidelberg, 2006.

[24]. Nichole, K. "HootSuite. (2011, May 10)." Measuring Social Media ROI: HootSuite White Paper Series (2011).

[25]. Tracker, LRA Crisis. "Resolve & invisible children." (2013).

[26]. Esuli, Andrea and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." In Proceedings of LREC, vol. 6, pp. 417-422. 2006.

[27]. Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417-424. Association for Computational Linguistics, 2002.

[28]. Han, Jiawei, Micheline Kamber and Jian Pei. *Data mining, Southeast Asia edition: Concepts and techniques*. Morgan kaufmann, 2006.

[29]. Alpaydin and Ethem. *Introduction to machine learning*. MIT press, 2014.

[30]. Kaur, Jasmeen and Vishal Gupta. "Effective approaches for extraction of keywords." *Journal of Computer Science* 7, no. 6 (2010): 144-148.

[31]. Masse, Mark. *REST API design rulebook*. “O’Reilly Media, Inc.", 2011.

[32]. Morstatter, Fred, Jürgen Pfeffer, Huan Liu and Kathleen M. Carley. "Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose." *arXiv preprint arXiv: 1306.5204* (2013).

[33]. Hardt, Dick. "The OAuth 2.0 authorization framework." (2012).

[34]. Gentry, Jeff and Maintainer Jeff Gentry. "Package ‘twitteR’." (2014).

[35]. Wickham, Hadley and Maintainer Hadley Wickham. "Package ‘plyr’." (2014).