Sentiment Analysis on Social Network/Media

Prepared By

Tejas Ruparelia

12MCEC22

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2014

Sentiment Analysis on Social Network/media

Major Project

Submitted in partial fulfillment of the requirements

For the degree of

Master of Technology in Computer Science and Engineering

Prepared By

Tejas Ruparelia

(12MCEC22)

Guided By

Prof Monika Shah

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2014

Certificate

This is to certify that the Major Project Report entitled "**Sentiment Analysis on Social Network/media**" submitted by **Tejas Ruparelia (Roll No: 12MCEC22)**, towards the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof Monika Shah
Guide & Associate Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof Vijay Ukani
Associate Professor
Coordinator M.Tech - CSE
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Undertaking for Originality of the Work

I, **Tejas Ruparelia**, Roll. No. **12MCEC22**, give undertaking that the Major Project entitled "**Sentiment Analysis on Social Network/media**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Nirma University, Ahmedabad, is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student
Date:
Place:

Endorsed by
Prof Monika Shah
(Signature of Guide)

Acknowledgements

It gives me a great pleasure in expressing thanks and profound gratitude to **Prof Monika Shah**, Department of Computer Engineering, Institute of Technology, Nirma University of Science and Technology, Ahmedabad for her great valuable guidance and continual encouragement throughout the project. I am heartily thankful to her for her time to time suggestion and clarity of the concepts of the topic that helped me a lot during this study. I am also very thankful to CSE deparment for provisioning of resources for the project and allowing me to work on such technologically advanced project.

My deepest thank you is extended to **Prof Vijay Ukani**, PG CSE - Coordinator, Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad for an exceptional support and continual encouragement throughout the Major Project.
It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.
A special thank you is expressed wholeheartedly to **Dr K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

The blessings of God and family members make the way for completion of Project. I am very much grateful to them.

- Tejas Ruparelia
12MCEC22

Abstract

Sentiment analysis is a quickly developing point for different requisitions, from promote world to inclining surveys. The basic trend was that the people get review from their relatives and companions before purchasing Product/Services, but today the trend is to recognize the opinion from different people whole around the world utilizing micro blogging information.

This dissertation work proposes to mine the sentiment analysis from a well known micro blogging services such as Twitter, where clients post their tweets or say their supposition(opinion) in regards to just about everything. In this framework it propose an approach in which a series of messages (tweets) from the Twitter micro blogging services is obtained and preprocessed them and then characterized focused around their emotional substance as neutral, polar(negative/positive) and irrelevant and dissects the interpretation of different classifiers are obtained in the form of Precisions and Recall.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General

In dislike past twitter has turned into a prevalent micro blogging administration that has a vast and quickly developing client where clients can compose their message which is known as ”tweets”. Clients utilize these tweets not just with respect to overhauling whatever is there in their brain, additionally to gives their supposition towards distinctive items, administrations, occasions and different things in which they are intrigued by. It has been seen that by and large the sentiments communicated by Twitter clients are helpful in certifiable circumstances, for example, item, administration audits on restaurants, gadgets, inns and so forth. By performing an assessment dissection on these tweets, advertisers ought to have the capacity to focus general society recognition about their items and administrations and buyers can equipped to know ahead of time what alternate clients think about the item or administration they are intrigued by.

Subscribe and distribute informal organization is the other name of twitter which helps by gives controlled connections among distinctive clients and has completely passionate information over a wide set of clients and diverse themes. Therefore, we can say that for client mining the estimations and conclusions from Twitter will be extremely valuable for some provisions. The opinion investigation on Twitter might be advantageous to an extensive variety of certifiable provisions from the arrangement of publicize focused around patterns, to requisitions which gather emotions towards a specific subject. Clients can get the input for a specific item or administration in order to settle on firm choice on their buy.

The territory of Sentiment Analysis gives suppositions and partitions them into diverse categories like polar, irrelevant, and neutral. Till now the vast majority of the work identified with notion dissection has been carried out on exploring destinations. Audit destinations give the estimations of items or motion pictures and in this manner confining the space of requisition to singularly business. Assumption dissection on Twitter posts is the following venture in the field of conclusion examination where it gives wealthier and more changed asset of presumptions that might be about anything from the most recent telephone they purchased, film viewed, political issues, religious perspectives or the people state of brain.

1.2 Objective of the Work

The fundamental goal of this exploration is to dissect different classifiers and to distinguish the viability and more competent classifier(s) for social networking (twitter) which could straight forwardness the procedure of arranging feelings in tweets. It might concentrate on recognizing distinctive classifiers with a satisfactory performance that could be utilized to arrange tweets focused around the communicated feeling as neutral, polar (negative or positive) and irrelevant.

1.3 Motivation of the Work

Because of the restriction of 140 characters in twitter individuals dependably utilize short structures and shortenings which could accumulate distinctive understandings diverse connections. Further the utilization of slang dialect and sentences with doubtful syntax expands the necessity for preprocessing. Because of this sort of restrictions in common dialect preparing this issue turns into a bottleneck in expanding the precision of the results. Consequently this territory is pulled in by diverse investigates which concentrate on enhancing the exactness and execution of distinctive characterization procedures.

1.4 Scope

This examination will give a methodology in which a series of messages (tweets) from the Twitter micro blogging sevices is obtained and preprocessed them and then characterized focused around their emotional substance as neutral, polar(negative/positive) and irrelevant and dissects the interpretation of different classifiers are obtained in the form of Precisions and Recall.

Chapter 2

Related Work

2.1 Related Work

Micro blogging services are rapidly growing new trend on the Internet. Traditionally blogs have been long posts which take many minutes to write but the Micro blogging enables users to post short or write information within moments. Thus, micro blogging platforms such as Twitter are a great way to discover what people know and give their opinion and communicate. So, discovering Twitter trends and its opinions has become a very popular field for research. A lot of research has been done in this field by researchers and scholars all around the world. Sentiment analyses in Tweets are typically done in two phases: (a) identifying sentiment expressions and (b) determining the polarity of the sentiment expressed in tweets. There have been different approaches used by researchers to classify tweets and analyze sentiments and trends in Twitter. Most researchers use lexical resources and decide the sentimentality of Tweets by the presence of lexical items [7]. Some other researchers combine additional features such as conjunction rules with lexical analysis to obtain more results with better accuracy [9].

One of the early researches in this field was done [8] where the authors try a novel approach to automatically classify sentiments in tweets. They have used distant learning methods where they classify tweets ending with positive emoticons such as : -) as positive; and tweets ending with negative emoticons such as : -(as negative. For feature space they try a unigram and bigram model. Their results indicate that the unigram model outperforms the bigram model. One drawback of this approach is that the data collected for this testing is through search queries in Twitter, which may be biased. The research also uses POS (Parts of speech) tags as features. The results of [10] have shown that using POS tags is not useful in classifying Tweets.

Another approach in sentiment classification in Twitter data where they label 1000 Tweets using polarity predictions from three different websites and another 1000 Tweets for testing. They explore some characteristics of tweets such as how they are written, as well as meta-information of the words used to compose them. In addition to polarity of words and POS of words they use syntax features of tweets such as re-tweets, hash tags, punctuations and exclamation marks. The accuracy of the test results obtained was very high due to (a) creating a more abstract representation of these messages instead of using raw word representation; and (b) labels of reasonable quality provided by data sources being combined.

Another popular research topic is to identify sentiments of popular brands such as Apple, Armani etc. This has a commercial value for the companies in question as they can identify the drawbacks of their products and improvements could be carried out as needed in order to be competitive. [6] use a commercial sentimental analyzer to analyze the sentiments of major brands. There have also been approaches to match opinion keywords [5]. For example, the hash tag sarcasm can be used to identify sarcastic Tweets about a brand or an event.

Chapter 3

Literature Survey

3.1 Data

The proposed approach required the use of different size of datasets. The data are collected from Twitter which required more effort than expected and it also require manual labeling of the posts for sentiments in relation to a query. However, the data which are collected is imbalanced and hence some sampling techniques or boosting techniques has to be applied for reducing the skewness of the data.

For doing sentiment analysis prepared the dataset with neutral,polar and irrelavent tweets which was physically gathered by questioning the Twitter Application Programming Interface (API) as there are just a couple of freely accessible datasets for tweets. These information were discretionarily browsed distinctive areas to guarantee mixture of information. There were no confinements on dialect or area was made throughout the gathering methodology of gathering this information. Each of the tweet is named as neutral, polar and irrelevant. The irrelevant marked tweets implies Non-English tweets.

Most of tweets on twitter are situated as open and it might be seen by any individual paying little respect to part of the twitter. The tweets which are termed as private are not seen by the general population. The information that is utilized as a part of this exploration was get from the general population course of events, and thus does not show the tweets that have been made confidential. Additionally the substance is gathered from the tweets for exploration and none of the personal information of the client are shown but only their opinion are shown.

3.2 Data Preprocessing

Because of the unusual nature of common dialect utilized as a part of twitter it is most probably that preprocessing strategies can be utilize for institutionalizing some amount of tweets. Most probably most of the tweets hold some manifestation of syntactic or spell check errors, acronym, sayings and jargon which is because of the 140 character impediment by Twitter on tweets.

Preprocessing procedure differentiates the pertinent substance from the tweets while forgetting the unimportant substance. Procedures connected are generally utilized as a part of data recovery requisitions particularly in opinion examination in micro-blogging administrations. Gathered information is passed through arrangement of distinctive preprocessors that is utilized as a part of the change for message thread into vector characteristic.

A portion of preprocessing steps which are done are clarified beneath. This is one of the crucial steps in whole grouping procedure as nature of properties that are concentrated from preparation dataset utilizing preprocessing procedure straightforwardly influences execution of algorithms.

- Replacing Emoticons

In numerous social networking blogs administrations, smileys are used by clients as a simple method for communicating the feeling. In this manner these emoticons are a simple approach to separate the non-polar messages from polar messages and negative from positive messages. These emoticons are supplanted with either a grin or annoyance keyword.

- Lower Case

For classifying the content it is extremely vital to have whole word in a reliable case in order to assurance that all words are guide to the similar characteristic independent of packaging. This for a greater extent important for this exploration work as it is extremely basic to discover spasmodic case in micro-blogging.

- Uppercase Identification

In numerous social networking blog administrations, it?s extremely normal to show the effective feelings in form of all capital letters. This practice we generally called as yelling and it could be viewed as a great indicator of the negative or positive of the message. So this preprocessing step extricates this characteristic of recognizing an arrangement of capital words before uprooting the packaging.

- URL Extraction

In numerous micro-blogging administrations numerous tweets hold URL in order to impart more substance as it has restricted character post. The substance in the URL may give supplementary learning with respect to the feeling a client attempting to express, in any case it might be exorbitant to creep the URL for their substance. In order to chop down the characteristic size throughout the preparation the whole URL in preparing tweets which has been supplanted with a class¡URL¿. It could respectably lessen the size.

- Detecting usernames and hash-tags

In tweets presents can point on alternate clients with utilization of @ image before a username and clients label the tweets relating to a category in twitter utilizing hash image. Again to keep away from difficulty of the characteristics, we supplant with a steady image. This substitution of usernames and hash labels diminish the characteristic size

- Punctuations Identifications

In numerous social networking blog administrations, its basic to utilize more punctuation within order to stay far from fitting syntax and to impart the feeling effectively. The punctuations can additionally give information about the positive or negative of the message. Case in point, shout imprints are utilized to express compelling articulation which is generally called as polar messages.

- Stop Words Removal

For classification, its a typical methodology to uproot stop words that are greatly normal which don't increase the value of the classification process. Basic words like an, an, and are on the whole known as stop words. As the consideration of using these words in the tweets do not provide any needed facts, so they are evacuated..

- Words Compression

Twitter clients have a tendency to be to utilize informal dialects and the greater part of them elaborate words to express the solid feelings. Case in point, the expression "saddddddddd" groups a higher degree statement than "miserable". Throughout the assessment and preparing for words holding more than three ensuing event of the same rehashed character we can diminish it to an arrangement of three character. For instance, we change over "sadddd" to "sadd" with the goal that the succession is not lessened to the more basic word character "dismal" which in order is utilization to separate between the general use and underscored use of particular character.

- Removal of Skewness in Dataset

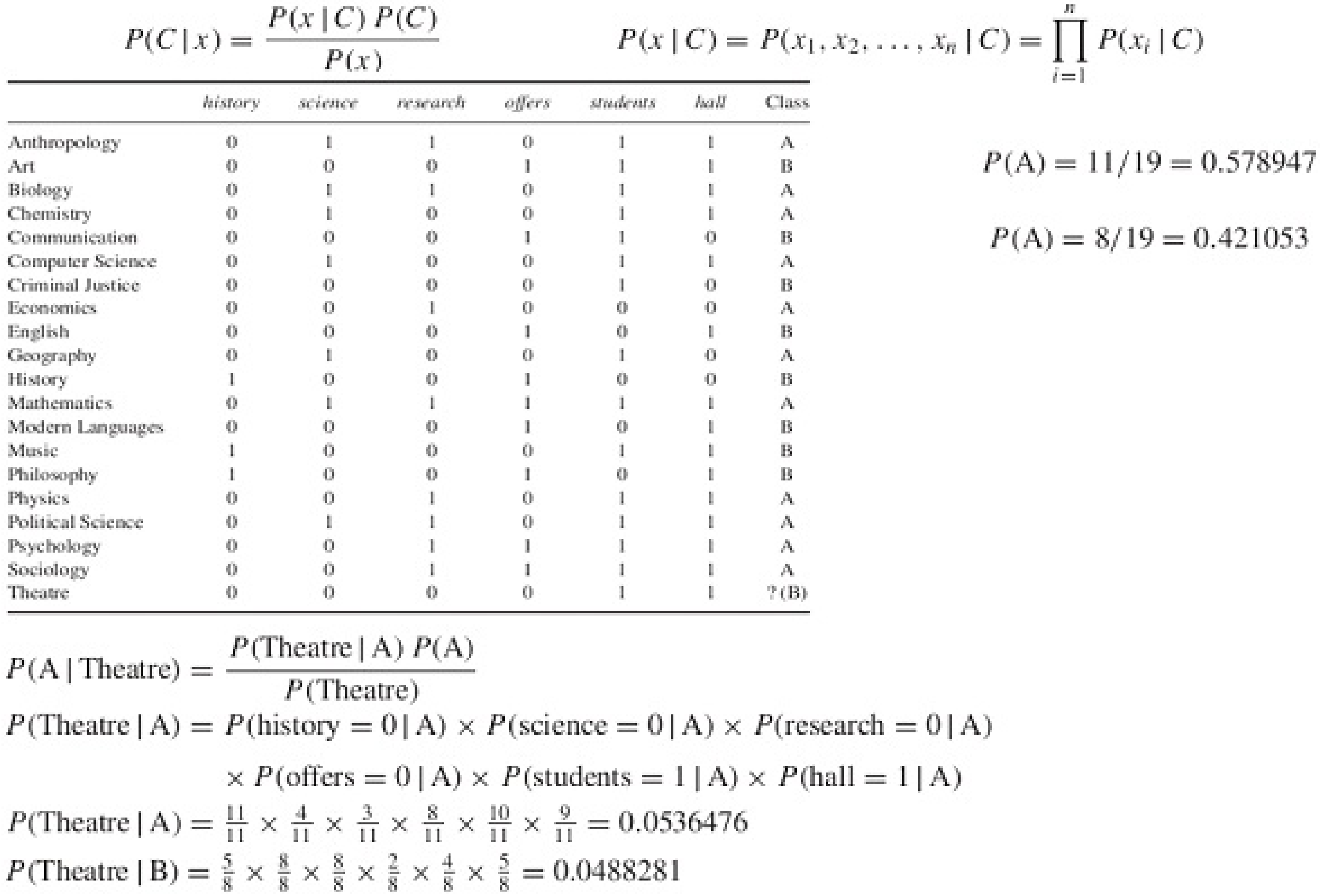
At the point when the preparation dataset is not adjusted, it is troublesome to building valuable classification models and which might be a testing undertaking. Class irregularity shows a issue in the utilization of essential classifier as they are endeavor to assemble models with the objective of accomplishing most extreme general classification exactness . Numerous procedures, for example, information testing and boosting strategies have been proposed to conquer the issues connected with irregularity class.

3.3 Various Classifiers

The sentiment analysis of the preprocessed twitter information has been completed utilizing diverse prominent classifiers which are talked about underneath, in order to recognize the good performance of the algorithms and separate their performance. The information mining and machine taking in apparatus Weka 3.6 version which was created and kept up by the University of Waikato, New Zealand has been utilized for classifying reason as it has various inherent classifiers with extra instruments of stopping the classifier. A percentage of the classifying algorithms that are regularly utilized as a part of studies are portrayed underneath:

- Naive Bayes

The Naive Bayes algorithm is a likely model focused around the Bayes' theorem, which figures the likelihood of tweet fitting in with a particular class, for example, positive, negative or neutral. This accepts that each characteristic are restrictively autonomous. Despite the fact that Naive Bayes classifier has yielded better brings about it didn't demonstrate better results contrasted with a bit different classifiers as delineated in such area.



$$P(C \mid x) = \frac{P(x \mid C)\,P(C)}{P(x)} \qquad P(x \mid C) = P(x_1, x_2, \ldots, x_n \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

	history	*science*	*research*	*offers*	*students*	*hall*	Class
Anthropology	0	1	1	0	1	1	A
Art	0	0	0	1	1	1	B
Biology	0	1	1	0	1	1	A
Chemistry	0	1	0	0	1	1	A
Communication	0	0	0	1	1	0	B
Computer Science	0	1	0	0	1	1	A
Criminal Justice	0	0	0	0	1	0	B
Economics	0	0	1	0	0	0	A
English	0	0	0	1	0	1	B
Geography	0	1	0	0	1	0	A
History	1	0	0	1	0	0	B
Mathematics	0	1	1	1	1	1	A
Modern Languages	0	0	0	1	0	1	B
Music	1	0	0	0	1	1	B
Philosophy	1	0	0	1	0	1	B
Physics	0	0	1	0	1	1	A
Political Science	0	1	1	0	1	1	A
Psychology	0	0	1	1	1	1	A
Sociology	0	0	1	1	1	1	A
Theatre	0	0	0	0	1	1	? (B)

$$P(\mathrm{A}) = 11/19 = 0.578947$$

$$P(\mathrm{A}) = 8/19 = 0.421053$$

$$P(\mathrm{A} \mid \mathrm{Theatre}) = \frac{P(\mathrm{Theatre} \mid \mathrm{A})\,P(\mathrm{A})}{P(\mathrm{Theatre})}$$

$$\begin{aligned} P(\mathrm{Theatre} \mid \mathrm{A}) = {} & P(\mathrm{history} = 0 \mid \mathrm{A}) \times P(\mathrm{science} = 0 \mid \mathrm{A}) \times P(\mathrm{research} = 0 \mid \mathrm{A}) \\ & \times P(\mathrm{offers} = 0 \mid \mathrm{A}) \times P(\mathrm{students} = 1 \mid \mathrm{A}) \times P(\mathrm{hall} = 1 \mid \mathrm{A}) \end{aligned}$$

$$P(\mathrm{Theatre} \mid \mathrm{A}) = \tfrac{11}{11} \times \tfrac{4}{11} \times \tfrac{3}{11} \times \tfrac{8}{11} \times \tfrac{10}{11} \times \tfrac{9}{11} = 0.0536476$$

$$P(\mathrm{Theatre} \mid \mathrm{B}) = \tfrac{5}{8} \times \tfrac{8}{8} \times \tfrac{8}{8} \times \tfrac{2}{8} \times \tfrac{4}{8} \times \tfrac{5}{8} = 0.0488281$$

Figure 3.1: Naive Bayes

- Random Forest

The majority of the choices rely on upon two information items created by random forests. At the point when the preparation set for the current tree is drawn by testing with substitution, about one-third of the cases are let alone for the example. This out-of-box information is utilized to get a running impartial evaluation of the order mistake as trees are added to the forest. It is likewise used to get appraisals of variable criticalness. After each one tree is assembled, the majority of the information are run down the tree, and vicinities are figured for each one sets of cases. On the off chance that two cases possess the same terminal hub, their vicinity is expanded by one. At the end of the run, the vicinities are standardized by separating by the amount of trees. Vicinities are utilized within supplanting missing information, finding outliers, and preparing enlightening low-dimensional perspectives of the information.

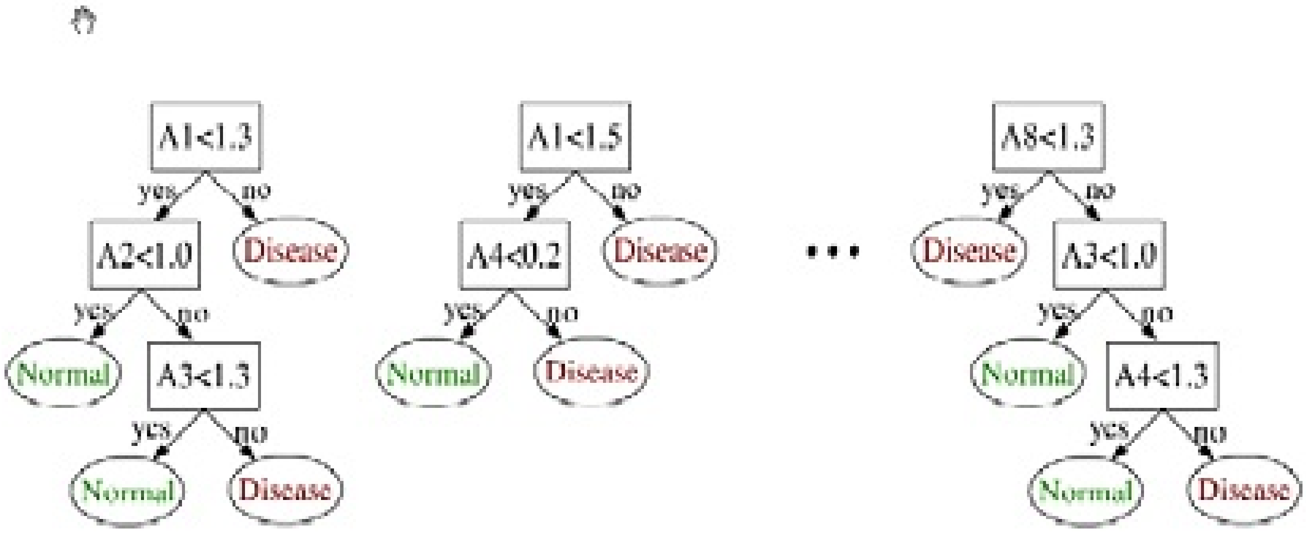


Figure 3.2: Random Forest

- Support Vector Machines (SVMs)

SVMs are the algorithms that are focused around piece substitution. They might be characterized as frameworks that utilize theory space of direct capacities in a tall dimensional characteristic space. This is prepared with a taking in algorithm that actualizes a taking in inclination inferred from factual taking in theory. It?s conceivable to build exceedingly non-linear classification system utilizing Support Vector Machines without getting stuck as a part of nearby minima.

Here we see the original articles (left half of the schematic) mapped, i.e., reworked, utilizing a set of scientific capacities, known as portions. The procedure of revamping the articles is known as mapping (transformation). Note that in this new setting, the mapped articles (right half of the schematic) is straightly distinct and, subsequently, as opposed to building the complex bend (left schematic), we should simply to discover an ideal line that can separate the GREEN and the RED items.

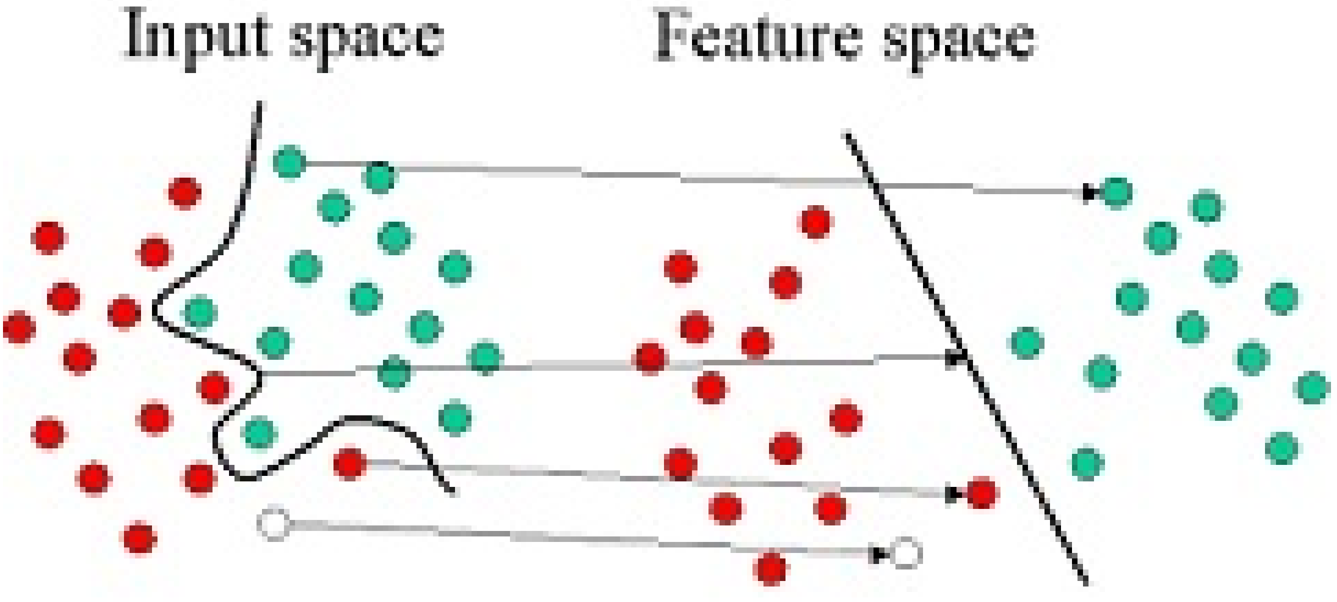


Figure 3.3: Support Vector Machines

- SMO

Sequential mining optimization algorithm proficiently tackles the optimisation issue when preparing support vector machines. SMO takes an iterative methodology to tackle the advancement issue where it breaks it into an arrangement of most diminutive conceivable sub-issues and explain them diagnostically. Every little issue includes two Lagrange multipliers as a result of the direct fairness constraint.the algorithm discovers a multiplier that maltreats KKT conditions and picks a second multiplier and upgrades the pair. It rehashes this methodology over and over until merging.

- J48 (C4.5 decision tree)

C4.5 is an algorithm created by Ross Quinlan to create choice trees from a set of preparing information. J48 is an open source java execution of C4.5 algorithm which is utilized within WEKA information mining device. The preparation information set is now characterized where each one specimen is a vector which speaks to the properties of the examples. C4.5 parts the example at every hub by picking a suitable trait of the information focused around information pick up.

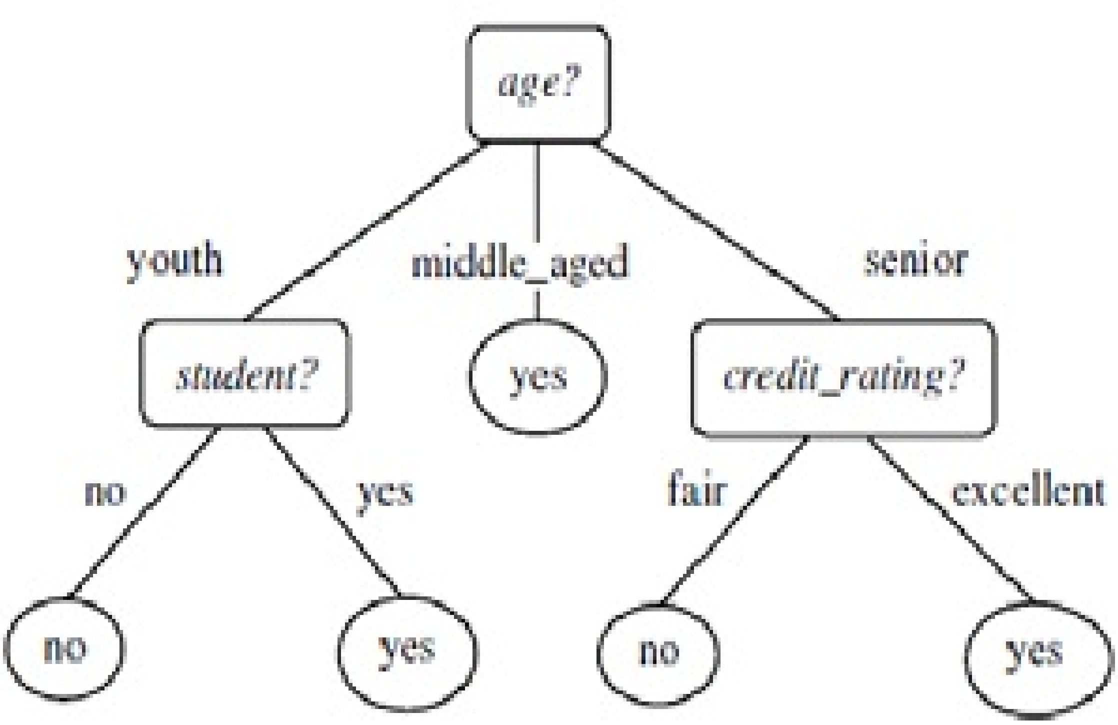


Figure 3.4: J48 (C4.5 decision tree)

Chapter 4

Proposed Work and Result Analysis

4.1 General Proposal

In this Proposed Experiment trying to develop an approach where a publicized stream of tweets from the twitter micro blogging site are preprocessed and classified based on their emotional content as neutral, polar (negative and positive) and irrelevant; and analyses the performance of various classifying algorithms based on their precision and recall in such cases.

When the training dataset is not balanced, it is difficult to building useful classification models and which can be a challenging task. Class imbalance presents a problem in the usage of basic classification algorithms as they are attempt to build models with the goal of achieving maximum overall classification accuracy . Many techniques such as data sampling and boosting techniques have been proposed to overcome the problems associated with imbalance class.

In this experiment, the above discussed classifiers were used for determining different size of datasets containing different size of instance and attribute. The training model was constructed by loading the preprocessed manually labeled tweets data into Weka 3.6 version and then using a cross validations method to train and test the classifier. Each test was carried out on a 3.50GHz Intel Core i7 based machine with 32GB of RAM, running on 64-bit version of the Microsoft Windows 7 operating system and 64-bit Oracle Java. The performance statistics for each of the classifier were captured. The time taken for training and testing of each run depended upon the nature of the classifier and the time taken typically varied from few minutes to more than six hours. For a dataset of size 5582 instances is used for neutral/polar/irrelevant classification and 982 for positive/negative classification.

4.2 Proposed System Architecture

- Following gives the idea of the classification process could be represent in a diagrammatic form as follows:

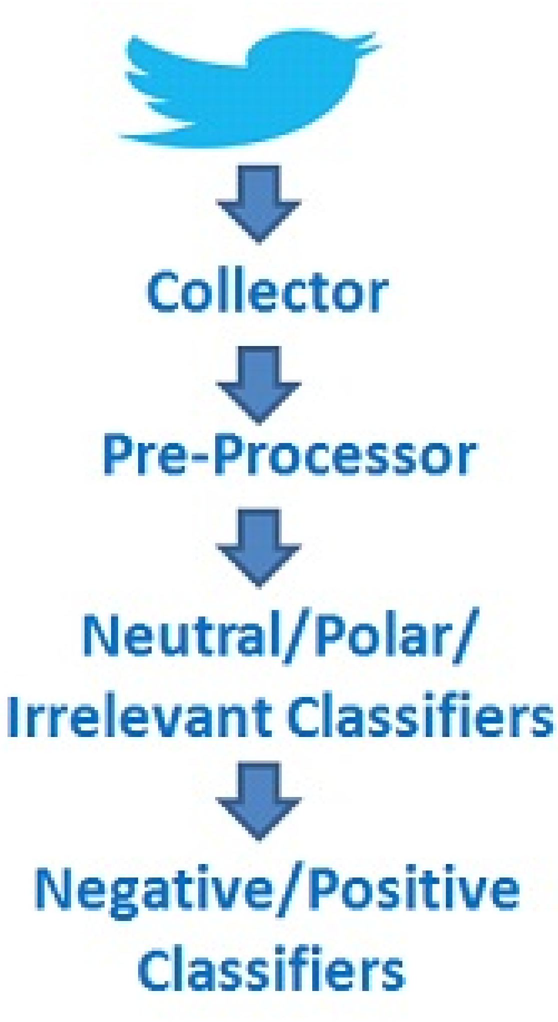
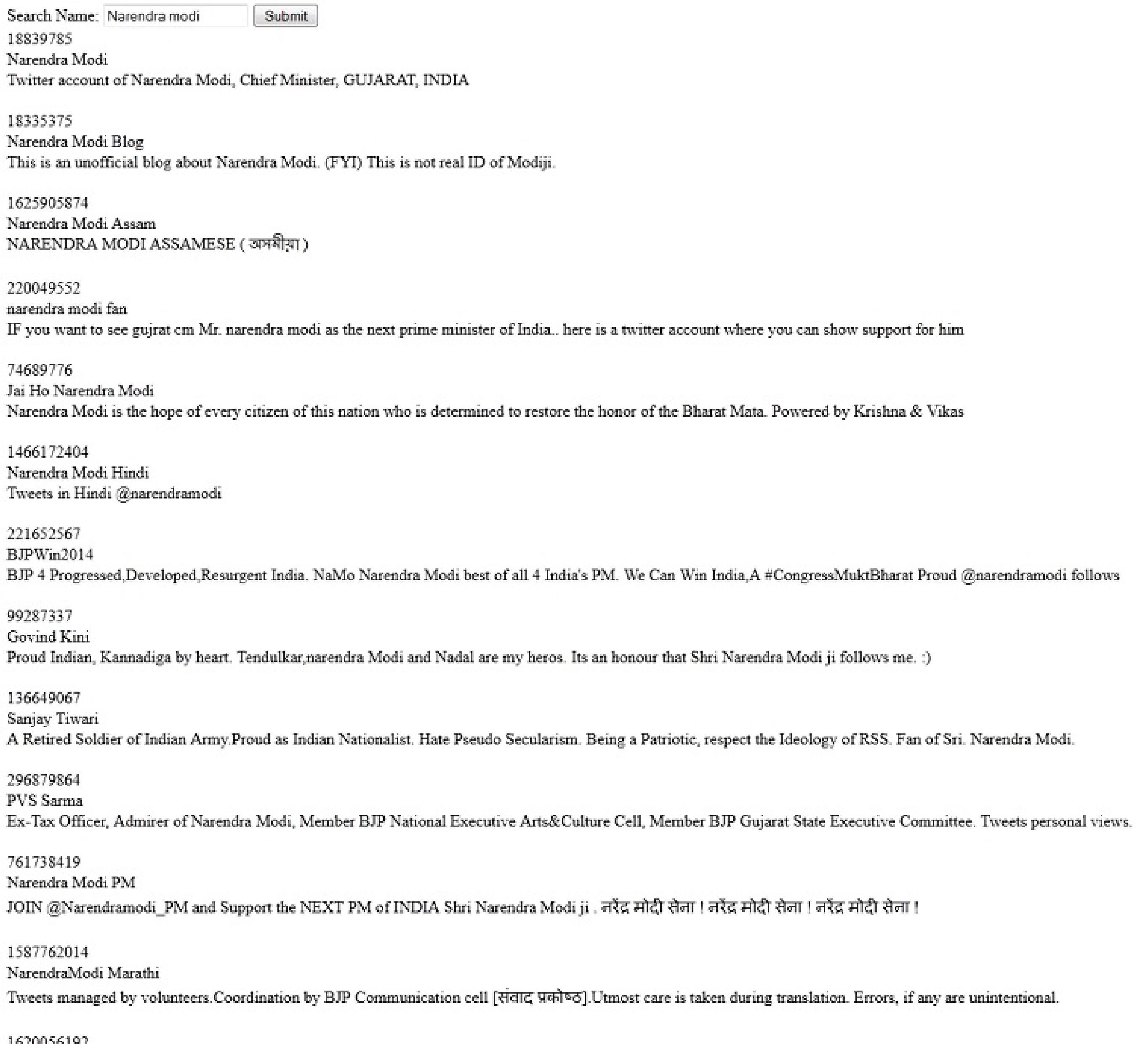


Figure 4.1: Proposed Process

4.3 Implementation Result and Discussion

Step 1:- Extraction of Tweets:

Twitter provide a twitter API which helps to extract the tweets from twitter.With the help of twitter API an application is devlop which use to extract the tweets according to the search keyword.

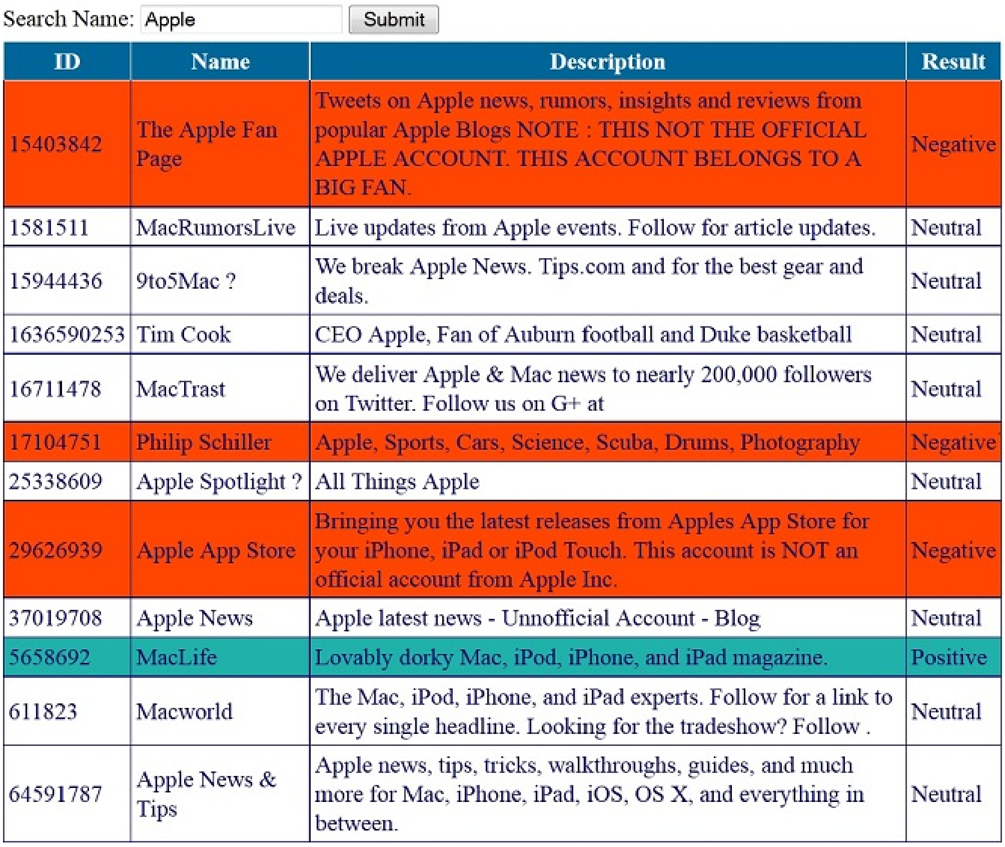


```
18839785
Narendra Modi
Twitter account of Narendra Modi, Chief Minister, GUJARAT, INDIA

18335375
Narendra Modi Blog
This is an unofficial blog about Narendra Modi. (FYI) This is not real ID of Modiji.

1625905874
Narendra Modi Assam
NARENDRA MODI ASSAMESE ( অসমীয়া )

220049552
narendra modi fan
IF you want to see gujrat cm Mr. narendra modi as the next prime minister of India.. here is a twitter account where you can show support for him

74689776
Jai Ho Narendra Modi
Narendra Modi is the hope of every citizen of this nation who is determined to restore the honor of the Bharat Mata. Powered by Krishna & Vikas

1466172404
Narendra Modi Hindi
Tweets in Hindi @narendramodi

221652567
BJPWin2014
BJP 4 Progressed,Developed,Resurgent India. NaMo Narendra Modi best of all 4 India's PM. We Can Win India,A #CongressMuktBharat Proud @narendramodi follows

99287337
Govind Kini
Proud Indian, Kannadiga by heart. Tendulkar,narendra Modi and Nadal are my heros. Its an honour that Shri Narendra Modi ji follows me. :)

136649067
Sanjay Tiwari
A Retired Soldier of Indian Army.Proud as Indian Nationalist. Hate Pseudo Secularism. Being a Patriotic, respect the Ideology of RSS. Fan of Sri. Narendra Modi.

296879864
PVS Sarma
Ex-Tax Officer, Admirer of Narendra Modi, Member BJP National Executive Arts&Culture Cell, Member BJP Gujarat State Executive Committee. Tweets personal views.

761738419
Narendra Modi PM
JOIN @Narendramodi_PM and Support the NEXT PM of INDIA Shri Narendra Modi ji . नरेंद्र मोदी सेना ! नरेंद्र मोदी सेना ! नरेंद्र मोदी सेना !

1587762014
NarendraModi Marathi
Tweets managed by volunteers.Coordination by BJP Communication cell [संवाद प्रकोष्ठ].Utmost care is taken during translation. Errors, if any are unintentional.

1620056192
```

Figure 4.2: Collection of Tweets

Step 2:- Preprocessing of tweets:

Preprocessing is one of the more vital steps in the entire classification process as the quality of the features/attributes that are extracted from the training dataset using the said preprocessing technique directly affects the performance of the classifiers. The preprocessing process extracts the relevant content from the tweets while leaving out the irrelevant ones. The techniques applied are used commonly in information retrieval applications specifically in sentiment analysis in micro-blogging. The collected data is passed through a series of pre-processors that assist in the conversion of the message strings into the feature vector.



Search Name: Apple Submit

ID	Name	Description	Result
15403842	The Apple Fan Page	Tweets on Apple news, rumors, insights and reviews from popular Apple Blogs NOTE : THIS NOT THE OFFICIAL APPLE ACCOUNT. THIS ACCOUNT BELONGS TO A BIG FAN.	Negative
1581511	MacRumorsLive	Live updates from Apple events. Follow for article updates.	Neutral
15944436	9to5Mac ?	We break Apple News. Tips.com and for the best gear and deals.	Neutral
1636590253	Tim Cook	CEO Apple, Fan of Auburn football and Duke basketball	Neutral
16711478	MacTrast	We deliver Apple & Mac news to nearly 200,000 followers on Twitter. Follow us on G+ at	Neutral
17104751	Philip Schiller	Apple, Sports, Cars, Science, Scuba, Drums, Photography	Negative
25338609	Apple Spotlight ?	All Things Apple	Neutral
29626939	Apple App Store	Bringing you the latest releases from Apples App Store for your iPhone, iPad or iPod Touch. This account is NOT an official account from Apple Inc.	Negative
37019708	Apple News	Apple latest news - Unnofficial Account - Blog	Neutral
5658692	MacLife	Lovably dorky Mac, iPod, iPhone, and iPad magazine.	Positive
611823	Macworld	The Mac, iPod, iPhone, and iPad experts. Follow for a link to every single headline. Looking for the tradeshow? Follow .	Neutral
64591787	Apple News & Tips	Apple news, tips, tricks, walkthroughs, guides, and much more for Mac, iPhone, iPad, iOS, OS X, and everything in between.	Neutral

Figure 4.3: Preprocessing of Tweets

Step 3:- Classification process:

In the next step the sentiment analysis of the preprocessed twitter data has been carried out using different popular classifiers which are discussed above, in order to identify the performance of the classifiers and differentiate their performance. The data mining and machine learning tool Weka which was developed and maintained by the University of Waikato, New Zealand has been used for classifying purpose as it has a number of built in classifiers with additional mechanisms of plugging the classifier.

- **Neutral/polar/irrelevant classification**

  - **Wordcount**

    Below table and graph shows the Accuracy and Recall of different classifier using wordcount information retrieval method for 5513 tweets

Classifiers	Accuracy	Avg F
Naive Bayes	58.84	0.599
Naive Bayes Multinomial	74.69	0.727
Complement Naive Bayes	74.17	0.742
DMNBtext	74.09	0.706
SMO	77.82	0.796
SVM	76.80	0.774
Random	70.82	0.712
J48	62.80	0.593
Lazy - TBK	68.80	0.692

Table 4.1: wordcount using 5513 tweets

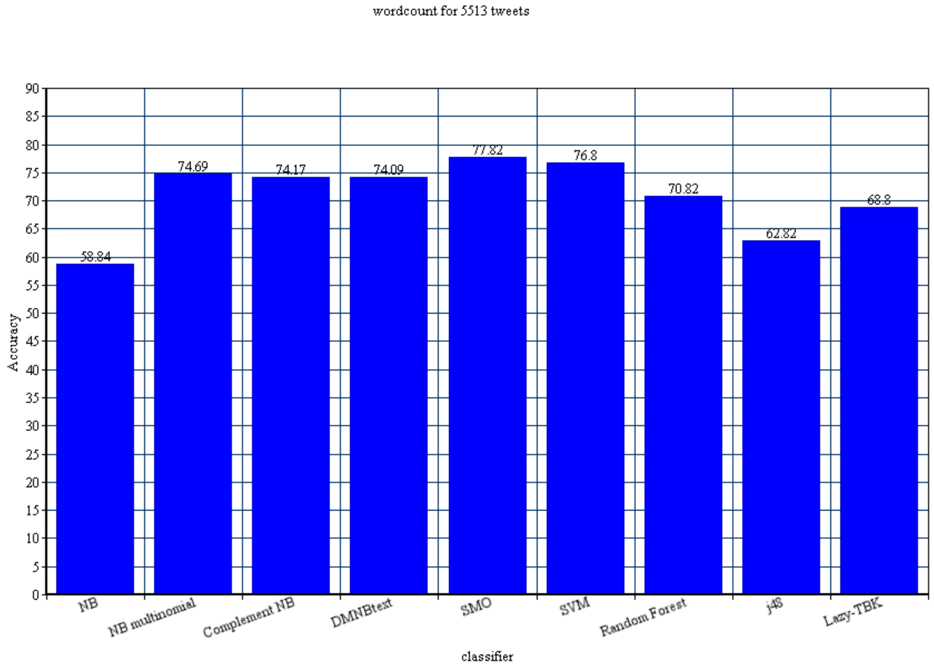


Figure 4.4: Wordcount using 5513 tweets

Below table and graph shows the Accuracy and Recall of different classifier using wordcount information retrieval method for 1500 tweets

Classifiers	Accuracy	Avg F
Naive Bayes	59.84	0.601
Naive Bayes Multinomial	75.69	0.737
Complement Naive Bayes	75.17	0.752
DMNBtext	75.09	0.716
SMO	79.82	0.799
SVM	78.80	0.782
Random	71.82	0.732
J48	63.83	0.622
Lazy - TBK	69.42	0.698

Table 4.2: wordcount using 1500 tweets

- **Graph View**

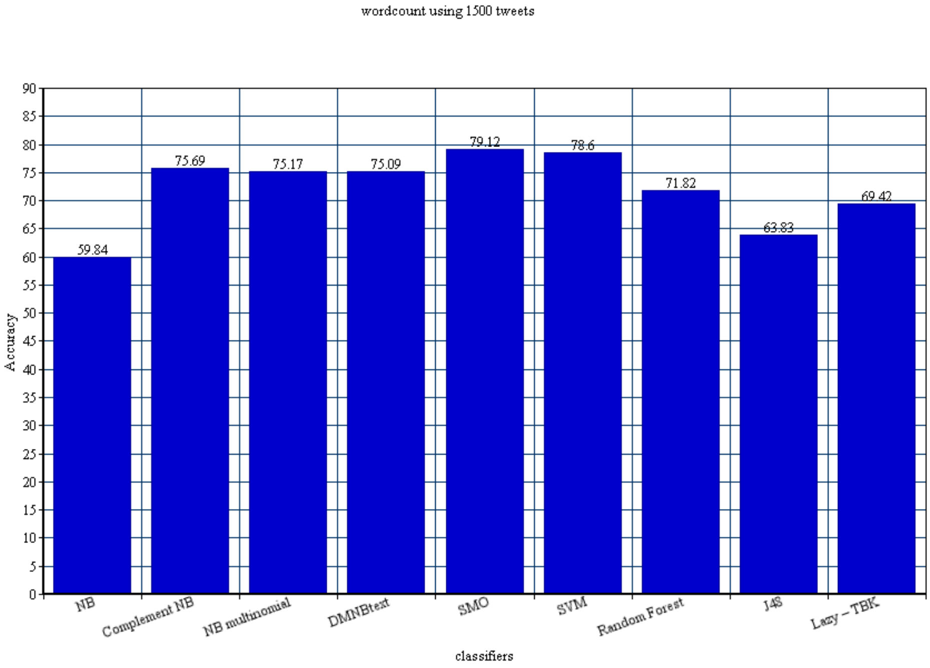


Figure 4.5: Wordcount using 1500 tweets

- **Boolean**

  Below table and graph shows the Accuracy and Rrecall of different classifier using Boolean information retrieval method for 5513 tweets

Classifiers	Accuracy	Avg F
Naive Bayes	58.84	0.599
Naive Bayes Multinomial	74.69	0.727
Complement Naive Bayes	74.17	0.742
DMNBtext	74.09	0.706
SMO	77.82	0.796
SVM	76.80	0.774
Random	70.82	0.712
J48	62.80	0.593
Lazy - TBK	68.80	0.692

Table 4.3: boolean using 5513 tweets

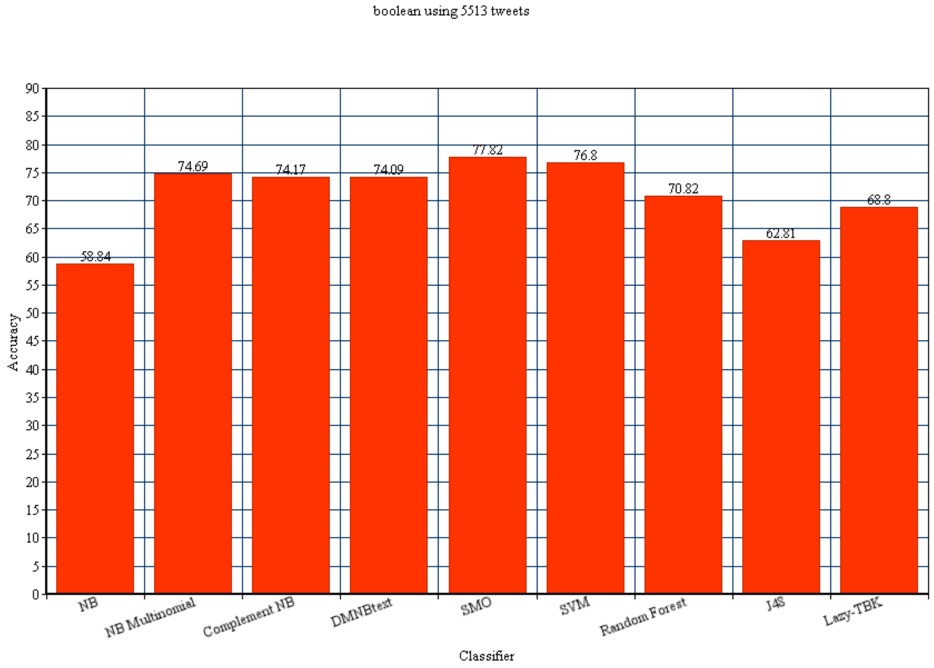


Figure 4.6: Boolean using 5513 tweets

Below table and graph shows the Accuracy and Recall of different classifier using Boolean information retrieval method for 1500 tweets.

Classifiers	Accuracy	Avg F
Naive Bayes	59.84	0.601
Naive Bayes Multinomial	75.69	0.737
Complement Naive Bayes	75.17	0.752
DMNBtext	75.09	0.716
SMO	79.82	0.799
SVM	78.60	0.782
Random	71.82	0.732
J48	63.83	0.622
Lazy - TBK	69.42	0.698

Table 4.4: boolean using 1500 tweets

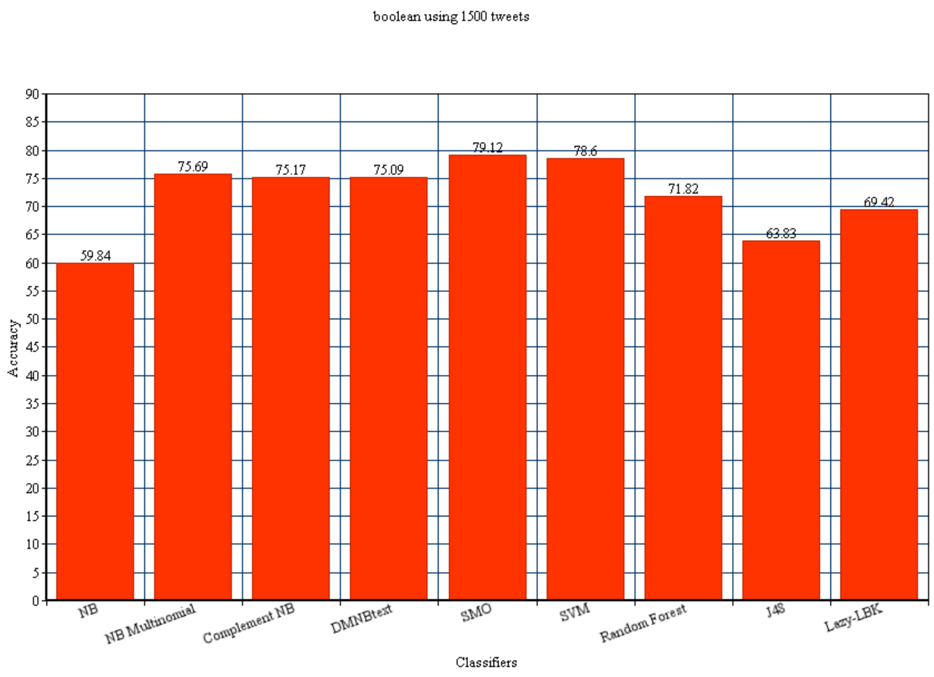


Figure 4.7: Boolean using 1500 tweets

- **TFIDF**

  Below table and graph shows the Accuracy and Recall of different classifier using TFIDF information retrieval method for 5513 tweets.

Classifiers	Accuracy	Avg F
Naive Bayes	58.93	0.597
Naive Bayes Multinomial	67.00	0.699
Complement Naive Bayes	72.46	0.734
DMNBtext	74.09	0.706
SMO	79.88	0.782
SVM	78.12	0.764
Random	70.11	0.701
J48	62.40	0.583
Lazy - TBK	67.80	0.682

Table 4.5: TFIDF using 5513 tweets

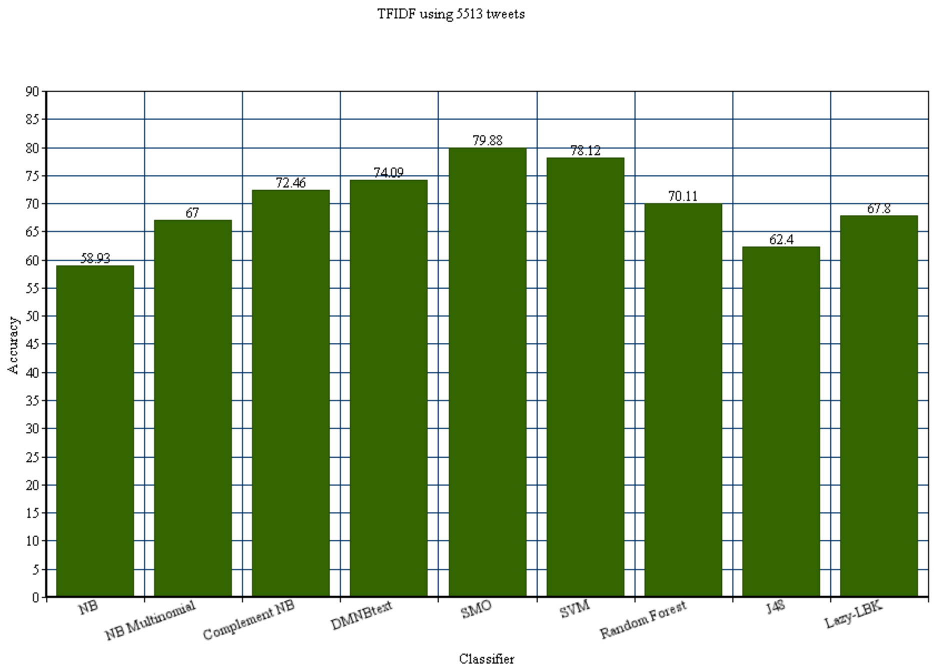


Figure 4.8: TFIDF using 5513 tweets

Below table and graph shows the Accuracy and Recall of different classifier using TFIDF information retrieval method for 1500 tweets.

Classifiers	Accuracy	Avg F
Naive Bayes	57.84	0.589
Naive Bayes Multinomial	74.69	0.727
Complement Naive Bayes	74.62	0.752
DMNBtext	74.20	0.720
SMO	79.32	0.782
SVM	77.80	0.772
Random	70.82	0.712
J48	62.83	0.620
Lazy - TBK	68.42	0.688

Table 4.6: TFIDF using 1500 tweets

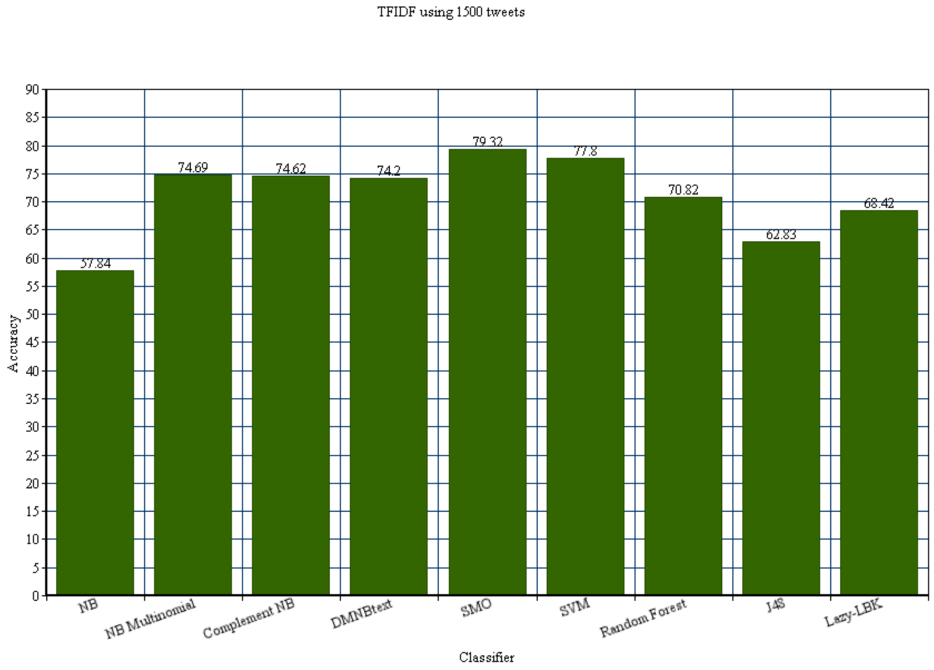


Figure 4.9: TFIDF using 1500 tweets

From the result it is seen that word count and Boolean contain almost same result whereas TFIDF give little less result and it was surprising to see that the widely used algorithms fail to perform satisfactory result, it is seen that simple naive bayes fail to gives the result while different bayes classifier perform better result than simple naive bayes classifier. The classifier SMO and SVM gives good acceptable result where as Lazy-IBK gives average result and classifier J48 gives poor result as compared to other classifier.

- **Positive/Negative classification**

  - **Wordcount**

    Below table and graph shows the Accuracy and Recall of different classifier using wordcount information retrieval method for positive/negative tweets.

Classifiers	Accuracy	Avg F
Naive Bayes	60.84	0.598
Naive Bayes Multinomial	76.69	0.747
Complement Naive Bayes	75.37	0.762
DMNBtext	74.19	0.719
SMO	81.82	0.798
SVM	80.14	0.785
Random	71.82	0.733
J48	63.80	0.593
Lazy - TBK	68.80	0.692

Table 4.7: Word Count using Positive/Negative Tweets

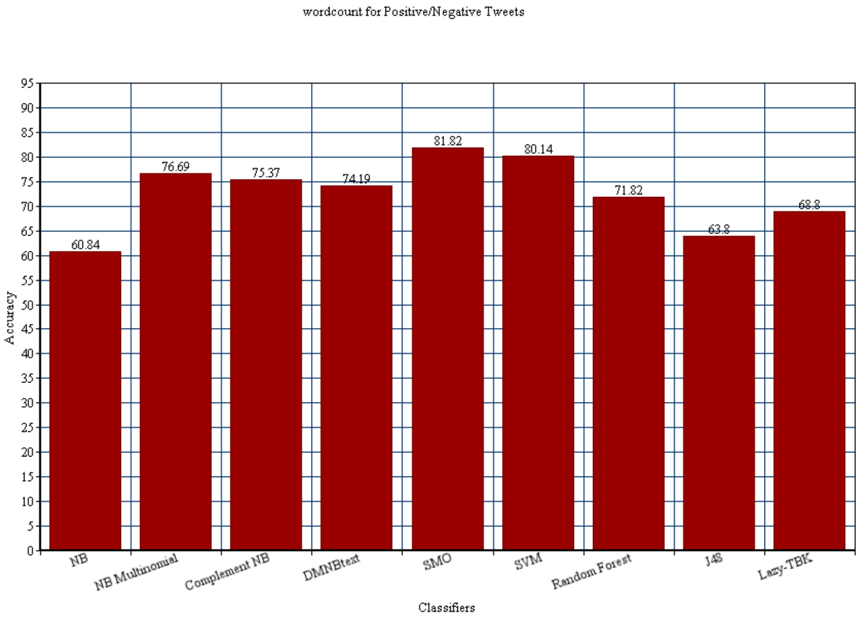


Figure 4.10: Wordcount using positive/negative tweets

- **Boolean**

  Below table and graph shows the Accuracy and Recall of different classifier using boolean information retrieval method for positive/negative tweets.

Classifiers	Accuracy	Avg F
Naive Bayes	60.84	0.598
Naive Bayes Multinomial	76.69	0.747
Complement Naive Bayes	75.37	0.762
DMNBtext	74.19	0.719
SMO	82.82	0.798
SVM	81.80	0.785
Random	71.82	0.733
J48	63.80	0.593
Lazy - TBK	68.80	0.692

Table 4.8: Boolean using Positive/Negative Tweets

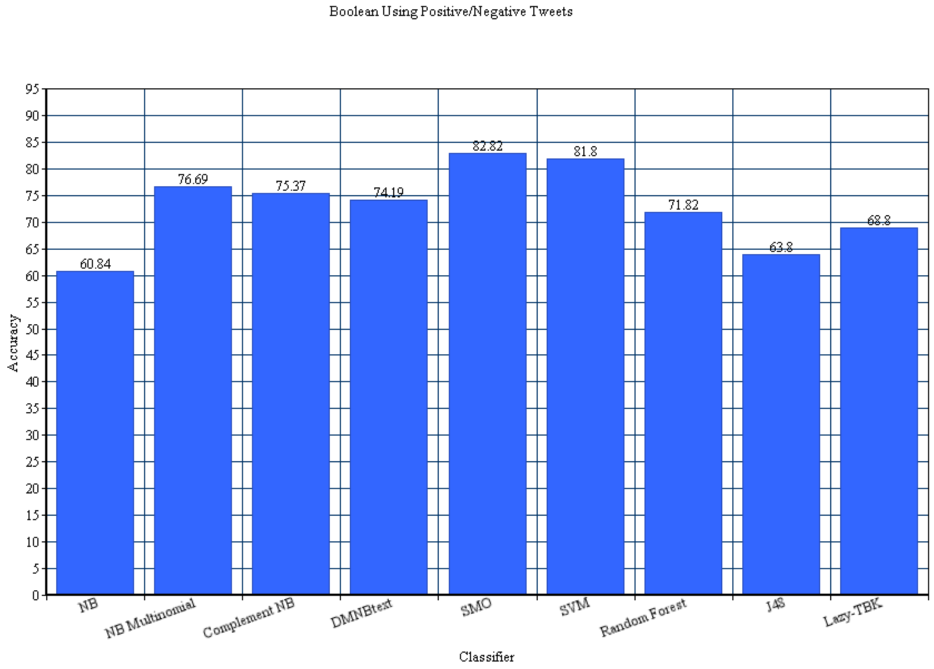


Figure 4.11: Boolean using positive/negative tweets

– **TFIDF**

Below table and graph shows the Accuracy and Recall of different classifier using TFIDF information retrieval method for positive/negative tweets.

Classifiers	Accuracy	Avg F
Naive Bayes	59.83	0.593
Naive Bayes Multinomial	76.01	0.761
Complement Naive Bayes	75.37	0.762
DMNBtext	74.44	0.744
SMO	81.89	0.782
SVM	80.99	0.788
Random	71.40	0.724
J48	63.45	0.612
Lazy - TBK	68.44	0.682

Table 4.9: TFIDF using Positive/Negative Tweets

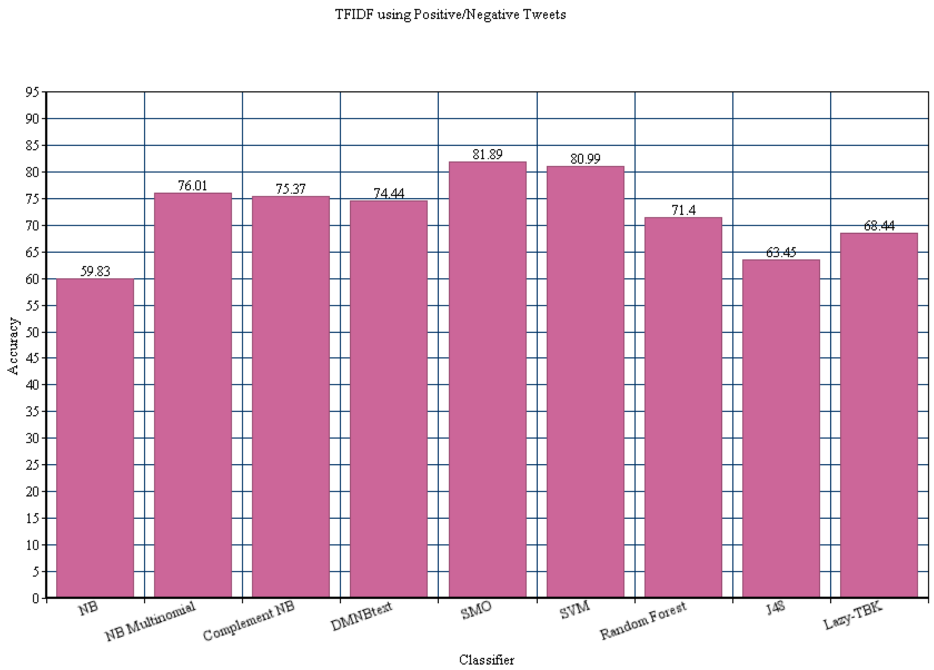


Figure 4.12: TFIDF using positive/negative tweets

From the result for positive/negative classification it was surprising to see that the widely used algorithms fail to perform satisfactory result, it is seen that simple naive bayes fail to gives the result while different bayes classifier perform better result than simple naive bayes classifier. The classifier SMO and SVM gives good acceptable result where as Lazy-IBK gives average result and classifier J48 gives poor result as compared to other classifier.

Chapter 5

Conclusion and Extension

From this study, it has been examine that the problem of sentiment analysis of Twitter micro blogging service is not quite the same as other classification issues. These twitter tweets are preprocessed using different dataset and then characterized focused around their emotional substance as neutral, polar(negative/positive) and dissects the interpretation of different classifiers are obtain in the form of Precisions and Recall.

From the result it is seen that wordcount and Boolean hold practically same result whereas TFIDF give minimal less result and it was astounding to see that the broadly utilized algorithms neglect to perform satisfactory result, as basic naive bayes neglect to gives good result while distinctive bayes classifier perform preferable over straight forward naive bayes classifier. The classifier SMO and SVM gives great satisfactory result where as Lazy-IBK gives acceptable result and classifier J48 gives poor result as contrasted with other classifier.

The sknewss of data set is still can be overcome to get better result and also some difficulties in natural language processing such as sarcasm detection etc could be overcome in future work.

Bibliography

[1] Balakrishnan Gokulakrishnan , Opinion Mining and Sentiment Analysis on a Twitter Data Stream, International Conference on Advances in ICT for Emerging Regions Dec 2012

[2] Akshi Kumar and Teeja Mary Sebastian, Sentiment Analysis on Twitter, IJCSI International Journal of Computer Science Issues July 2012

[3] F. Bhat, M. Oussalah, K. Challis and T. Schnier , A Software System for Data Mining with Twitter, 2011 10th IEEE International Conference-Sept 12

[4] Son Doan, Lucila Ohno-Machado, Enhancing Twitter Data Analysis with Simple Semantic Filtering, 2012 IEEE Second Conference on Healthcare Informatics

[5] Changhyun Byun and Hyeoncheol Lee, Automated Twitter Data Collecting Tool for Data Mining in Social Network, IJCSI International Journal of Computer Science Issues Dec 2012

[6] Book: Web Data Mining Exploring Hyperlinks, Contents, and Usage Data by Bing Liu.

[7] Book: Data Mining: Concepts and Techniques by Jiawei Han and Micheline Kamber.

[8] B. Pang and L. Lee, -Opinion Mining and Sentiment Analysis In Foundations and Trends in Information Retrieval, vol. 2, pp. 1-135, 2008.

[9] B. Pang, L. Lee, and S. Vaithyanathan, -Thumps up? Sentiment Classification Using Machine Learning Techniques, in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP),2002

[10] L. Barbosa and J. Feng, -Robust sentiment detection on twitter from biased and noisy data, in Proc. 23rd International Conference on Computational Linguistics: Posters, 2010

[11] S. Prasad, -Micro-blogging Sentiment Analysis Using Bayesian Classification Methods, Technical Report, 2010

[12] WordNet: http://wordnet.princeton.edu/

[13] twitter: https://dev.twitter.com/issues