

# **TWITTER BASED SENTIMENTAL ANALYSIS USING ENHANCE NAÏVE BAYES**

*Dissertation submitted in fulfilment of the requirements for the Degree of*

## **MASTER OF TECHNOLOGY**

**In**

## **COMPUTER SCIENCE AND ENGINEERING**

By

**AMIT KUMAR SINGH**

**Registration number**

**11500996**

Supervisor

**Mr. Robin Prakash Mathur**



**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

Month April, Year 2017

@ Copyright LOVELY PROFESSIONAL UNIVERSITY, Punjab (INDIA)

Month April, Year2017.

ALL RIGHTS RESERVED

## ABSTRACT

---

Information mining is a procedure of getting a concealed learning and pattern from the real-time information set. In information mining, the essential concern is the pre-handling of information for making the information possible for mining. Social media in general exhibit a rich variety of information sources: in addition to the content itself, there is a wide array of non-content information available, such as links between items and explicit quality ratings from members of the community. In our research work we proposed an algorithm which enhance the Naïve Bayes and increased the accuracy rate as well as try to predict the current situation on the basis of analysis the sentiments of the users who updated their thoughts on the social media like twitter on the current issues.

## DECLARATION STATEMENT

---

I hereby declare that the research work reported in the dissertation entitled **“TWITTER BASED SENTIMENTAL ANALYSIS USING ENHANCE NAÏVE BAYES”** in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. **Robin Prakash Mathur** I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

*Signature of Candidate*

**Amit Kumar Singh**

**Registration No. 11500996**

## SUPERVISOR'S CERTIFICATE

---

This is to certify that the work reported in the M.Tech Dissertation entitled **“TWITTER BASED SENTIMENTAL ANALYSIS USING ENHANCE NAÏVE BAYES”**, submitted by **Amit Kumar Singh** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Mr. Robin Prakash Mathur

**Date:**

**Counter Signed by:**

**1) Concerned HOD:**

HoD's Signature: \_\_\_\_\_

HoD Name: \_\_\_\_\_

Date: \_\_\_\_\_

**2) Neutral Examiners:**

**External Examiner**

Signature: \_\_\_\_\_

Name: \_\_\_\_\_

Affiliation: \_\_\_\_\_

Date: \_\_\_\_\_

**Internal Examiner**

Signature: \_\_\_\_\_

Name: \_\_\_\_\_

Date: \_\_\_\_\_

## ACKNOWLEDGEMENT

---

I have taken efforts in this dissertation-II. However it, would not have been possible without the kind support and help of my mentor **Mr. Robin Prakash Mathur**. I would like to extend my sincere thanks to him. I am highly indebted for his guidance and constant supervision as well as for providing necessary information regarding the dissertation and also for their support in completing the Dissertation-II.

I would like to express my gratitude towards **Er. Dalwinder Singh HOD (CSE)**, members of Lovely Professional University for their kind co-operation and encouragement which help me in completion of this dissertation.

My thanks and appreciation also go to my colleagues in the doing the Dissertation –II and people who have willingly helped me out with their abilities.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE NO.</b>
Cover Page	i
PAC form	ii
Abstract	iii
Declaration by the Scholar	iv
Supervisor's Certificate	v
Acknowledgement	vi
Table of Contents	vii
List of Figures	ix
<b>CHAPTER1: INTRODUCTION</b>	<b>1</b>
<b>1.1 SENTIMENTAL ANALYSIS</b>	<b>1</b>
<b>1.2 TYPES OF SENTIMENTAL ANALYSIS</b>	<b>2</b>
<b>1.3 DOCUMENT PRE-PROCESSING</b>	<b>5</b>
<b>CHAPTER2: REVIEW OF LITERATURE</b>	<b>7</b>
<b>CHAPTER3: PRESENT WORK</b>	<b>18</b>
<b>3.1 PROBLEM FORMULATION</b>	<b>18</b>
<b>3.2 OBJECTIVES OF THE STUDY</b>	<b>19</b>
<b>3.3 METHODOLOGY</b>	<b>19</b>
<b>3.3.1 RESEARCH METHODOLOGY</b>	<b>19</b>
<b>3.3.2 DEVELOPMENT TOOL</b>	<b>21</b>
<b>3.3.3 FLOW CHART</b>	<b>23</b>

## TABLE OF CONTENTS

CONTENTS	PAGE NO.
<b>CHPTER4: RESULTS AND DISCUSSION</b>	24
<b>4.1 EXPERIMENTAL RESULTS</b>	24
<b>4.2 COMPARISION WITH EXISTING TECHNIQUE</b>	32
<b>CHAPTER5: CONCLUSION AND FUTURE SCOPE</b>	36
<b>5.1 CONCLUSION</b>	36
<b>5.2 FUTURE SCOPE</b>	36
<b>REFERENCES</b>	37
<b>APPENDIX</b>	41



## LIST OF FIGURES

FIGURE NO.	FIGURE DESCRIPTION	PAGE NO.
<b>Figure1.1</b>	Block Diagram of Analysis	3
<b>Figure1.2</b>	Basic Algorithm Approach	4
<b>Figure1.3</b>	Flow Chart of Data Pre-Processing	5
<b>Figure2.1</b>	Twitter Monitor Architecture	8
<b>Figure2.2</b>	System Used	14
<b>Figure3.1</b>	Why R for Analysis	22
<b>Figure3.2</b>	Flow Diagram of Research Methodology	23
<b>Figure4.1</b>	Proposed System	24
<b>Figure4.2</b>	Tweets Fetch From Twitter Using R	25
<b>Figure4.3</b>	Fetching Tweets	26
<b>Figure4.4</b>	Mean of Top 20 Words	27
<b>Figure4.5</b>	Probability of words	28
<b>Figure4.6</b>	Word Cloud	29
<b>Figure4.7</b>	Frequency of Top 10 Words	30
<b>Figure4.8</b>	Accuracy of Enhance Naïve Bayes	31
<b>Figure4.9</b>	Accuracy of Naïve Bayes	31
<b>Figure4.10</b>	Precision Comparison	32
<b>Figure4.11</b>	TPR Comparison	33
<b>Figure4.12</b>	Accuracy Comparison of Different Classifiers	34
<b>Figure4.13</b>	Comparative Analysis	35

## Checklist for Dissertation-II Supervisor

Name: \_\_\_\_\_ UID: \_\_\_\_\_ Domain: \_\_\_\_\_

Registration No: \_\_\_\_\_ Name of student: \_\_\_\_\_

Title of Dissertation:

\_\_\_\_\_

- 
- ☐ Front pages are as per the format.
  - ☐ Topic on the PAC form and title page are same.
  - ☐ Front page numbers are in roman and for report, it is like 1, 2, 3.....
  - ☐ TOC, List of Figures, etc. are matching with the actual page numbers in the report.
  - ☐ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.
  - ☐ Color prints are used for images and implementation snapshots.
  - ☐ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.
  - ☐ All the equations used in the report are numbered.
  - ☐ Citations are provided for all the references.
  - ☐ **Objectives are clearly defined.**
  - ☐ Minimum total number of pages of report is 50.
  - ☐ Minimum references in report are 30.

Here by, I declare that I had verified the above mentioned points in the final dissertation report.

Signature of Supervisor with UID

# CHAPTER 1

## INTRODUCTION

---

Information mining is a procedure of getting a concealed learning and pattern from the real-time information set. In information mining, the essential concern is the pre-handling of information for making the information possible for mining. The path toward cleaning the data is in like manner called as uproar transfer or noise reduction [1]. Information happens to be the key worry in information mining with the gigantic online information created from a few sensors, Web Hand-off Visits, Twitter, confront book, Online Bank or ATM. Exchanges, the idea of powerfully changing information is turning into a key test, what we call as information streams. Twitter is a miniaturized scale blogging administration that checks with a huge number of clients from everywhere throughout the world. It permits clients to post and trade 140-character-long messages, which are otherwise called tweets. Twitter is utilized through a wide assortment of customers, from which a substantial segment – 46% of dynamic clients – compare to portable clients. Tweets can be distributed by sending messages, sending SMS instant messages and straightforwardly from cell phones utilizing a wide cluster of Web-based administrations [2]. Thusly, twitter encourages the ongoing spread of data to an expansive gathering of clients. This makes it a perfect domain for the dispersal of breaking-news specifically from the news source and/or topographical area of occasions.

### **1.1 What is Sentiment Analysis?**

The sentimental analysis is a process which uses natural language processing, the concept of machine learning methods to get, identify and characterize the content of sentiments into a text unit. It referred as opinion mining, although its emphasis, in this case, is on extraction [4]. In simple word, we define the wistful analysis as the procedure of computationally recognizing and classifying conclusions communicated in a type of content particularly keeping in mind the end goal to decide the author's state of mind towards a specific subject, item, and so on i.e. positive, negative, or unbiased.

## 1.2 Types of sentimental Analysis

**Objectivity/Subjectivity Identification:** -The term defines as objective/subjective identification is commonly defined as classifying a given text (usually a sentence) into one of two classes i.e. objective or subjective [4]. The issue with target/subjective recognizable proof is that it can some of the time be more troublesome than extremity arrangement. The subjectivity of words and expressions may rely on upon their unique circumstance and a target archive may contain the subjective sentences, for example, a news article citing individuals' suppositions and so on.

**Feature/aspect-based:** The term refers to deciding the conclusions or notions which can be communicated to various components or parts of substances, e.g., of a wireless, a computerized camera, or banks. An element/viewpoint is a quality or part from an element, e.g. those screens of a Mobile telephone, the organization to an eatery, or the photo way of a Polaroid. [4] The advantage of this type of sentiment analysis is to gain the capture nuances about objects of interest. Sentiment responses may be different because of different feature generate different sentiments, for example, a hotel can have a convenient location, but mediocre food This issue includes a few sub-issues, e.g., distinguishing important substances, separating their elements/viewpoints, and figuring out if a conclusion communicated on every element/angle is sure, negative or impartial The programme ID of elements can be performed with syntactic techniques or with subject displaying [6].

The sentimental analysis is a set of collection of methods which is generally implemented in computer science to study the behavior of the society based on the current change occurs in the country such as government decision, economy, trade etc. It is a broad field of study because it involves the behavior, thoughts, nature, opinion of the user on that basis the mind of a people is very easily read by this analysis. As the sentimental investigation has enhanced in the last few decades so has its applications. Opinion/sentiment analysis is currently being utilized from particular item showcasing to hostile to social conduct acknowledgment [18].

The propels in all the social media site and other microblogging and long range interpersonal communication locales have not just contributed to change to social locales, however, on a very basic level changed a way that we utilize these

destinations what's more, that how we share our sentiments. As opinion mining has enhanced in the last few decades so has its applications. The wistful analysis is currently being utilized from particular item showcasing to hostile to social conduct acknowledgment [15].

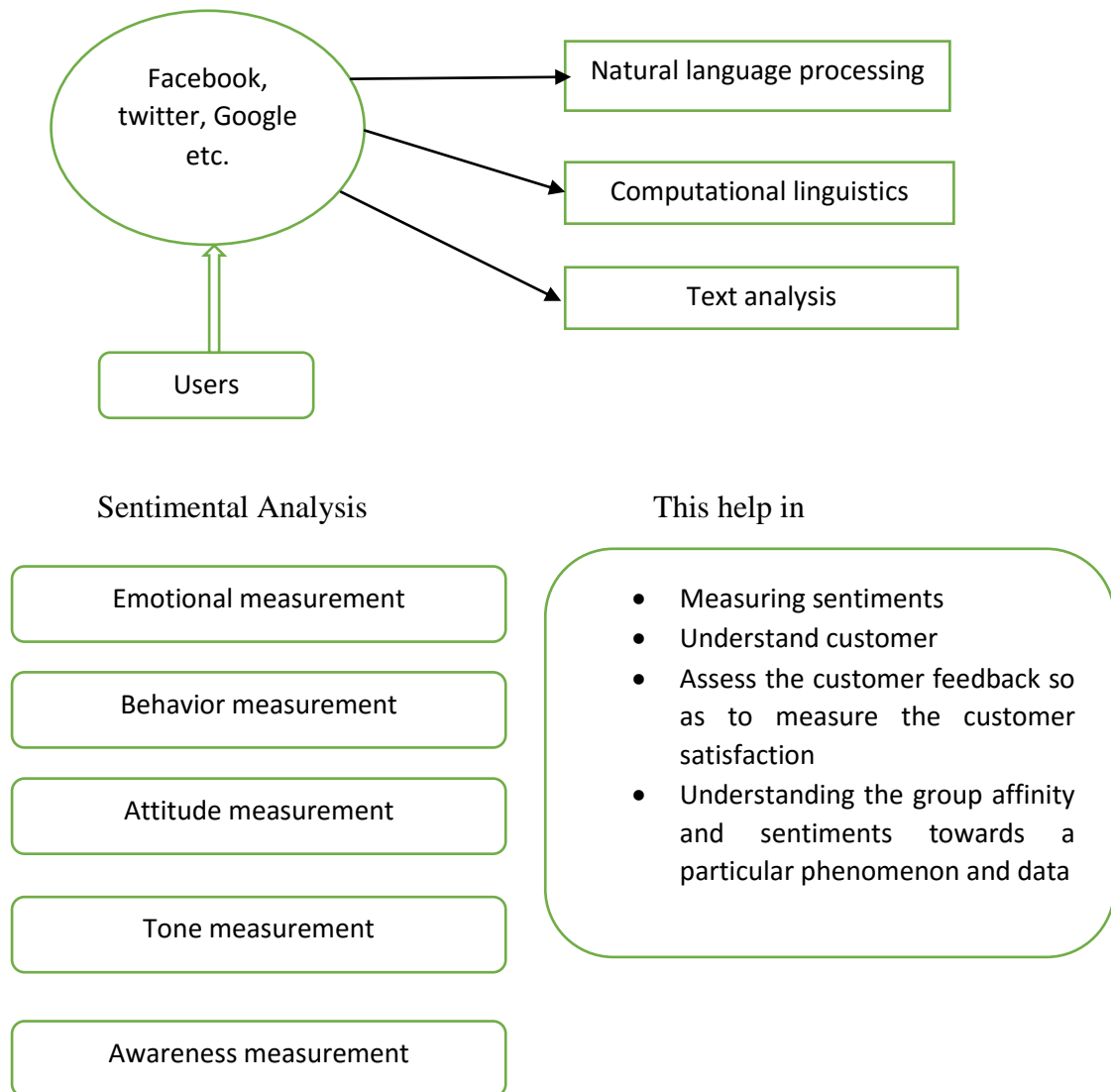


Figure 1.1: Block Diagram of analysis

The propels in Facebook, twitter, YouTube and different microblogging. Millions of messages seen day by day in well-known sites that give administrations to microblogging [11]. Writers of those messages compose about their life, impart insights on an assortment of points and examine current issues. In view an unrestricted organization of posts and a simple accessibility of microblogging stages, Internet consumers tend to move from customary specialized apparatuses to

microblogging administrations [5]. As to an ever increasing extent clients post about items and administrations they utilize or express their political and religious perspectives, microblogging sites get to be important wellsprings of people's assessments and estimations. Such data can be beneficially used for flexible UI and long range interpersonal correspondence goals have not quite recently contributed change to the social goals yet have on an exceptionally essential level changed the way we use these regions in addition, how we share our feelings, our viewpoints with the broader gathering of spectators[14]. A fundamental assignment in assumption examination is characterizing the extremity of a given content at the record, sentence, or highlight/perspective level whether the communicated sentiments in an archive, a sentence or an element include/angle is sure, negative, or impartial. Progressed, "past extremity" supposition characterization looks, for example, at passionate states, for example, "furious", "pitiful", and "cheerful" [8]. The sentimental examination can improve the client encounter over an interpersonal organization or framework interface. The learning calculation will take in what our feelings are from factual information then decide the state of mind. After that, it will change our social collaborations as needs are on our interpersonal organization destinations or different interfaces like desktop or framework administrations or website pages [3].

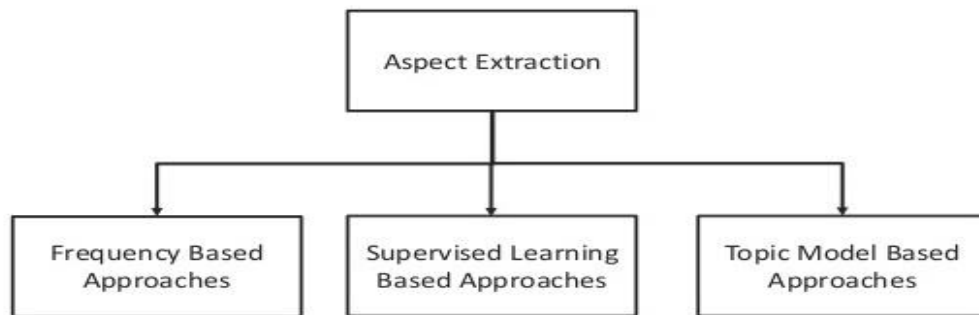


Figure 1.2: Basics Algorithms Approach

Text mining is a technique which is used for finding the information from a text-based databank. It is a process of mining interesting facts or information from amorphous textual database or documents. The textual database includes the vast collection of text forms from different sources such as digital libraries, social media, education websites, e-commerce sites, mail services etc.

### 1.3 DOCUMENTS PRE-PROCESSING PROCEDURE

Pre-processing is a method in which we can handle our data or prepare our data for the mining process. Information mining has also gone through the pre-processing method called ETL (extract, transform and load) same as in the text mining we also use pre-processing of the data for the accurate and effective result.

In the text mining the pre-processing of data can be done into three phase which is as follows:

- Tokenization
- Stop words removal
- Stemming

Tokenization has recognized the keywords for expressive the forms. In other words, it is a process of severe the sentence into many distinct tokens. Such as ram is always online is split as ram/is/always/online.

Stop words removal is another important step of data pre-processing which means to remove the unwanted or unnecessary words from the data. In another word, the stop words removal is the technique which is used to remove the unwanted words, for example, can, is, was, there, etc.

The final phase of data pre-processing method is stemming which worth multiple words may share the same meaning with same word stem. Such as the word love, loveable, loved, loving are showing the same stem word love. In other words, we say that stemming is the process of removing the suffix words from the stem word.



Figure 1.3: Flow chart of data pre-processing

The above figure 1.3 explain the process of data pre-processing, in which the raw data (tweets) which contain lots of unusual words is fetched by the twitter API for the analysis task but due to unusual words the accuracy and performance of the system is compromised due to which we perform the pre-processing task to filter the data. As the

flow chart explain the first step is tokenization, the second step is to stop word is removed the third step is to remove stemming and after stemming we get the processed data which is suitable for the text mining and for analysis task.



## CHAPTER 2

### REVIEW OF LITERATURE

---

C. Castillo et al. [19] proposed the methodology to check or analysis the credibility of news propagated through twitter. The main target of this is to check the credibility of data spread through social media network. They first collect the data (tweets) from the twitter using twitter API then after they twitter monitor to detect the tweets of about a 2-months period. Twitter Monitor is the on-line monitoring software system which detects the increases (“bursts”) in the frequency of keywords found in messages. They collected every tweet which matches the query under a 2-day window centered on the peak of every burst. Newsworthy theme assessment was their first marking round was expected to isolate the subjects who spread data about a news occasion, from the cases which relate to individual suppositions and visit. Over different word, they separated the messages that need aid for possibility premium to an expansive situated for people, starting with discussions that are of minimal significance outside for this task we used Mechanical Turk8.

Next, they focus on credibility assessment for this they use the classification of the tweets which was gathered by the API and by applying the algorithm, they analysis whether the tweets related to information/news or belongs from the personal chat. On this paper, the author tries to separate the automatically separate the newsworthy topic from the other type of conversation from time-sensitive data.

M. Mathioudakis et al. [13] explain the use of twitter monitor (an online monitoring system) which detect the frequent keywords from the tweets. The frequent words which were detected by the system known as a trend. It is also important for the marketing professionals and for opinion tracking companies, the main challenges for the researcher from their point of view are automatically detected and analysis the hot topics (trend). According to this paper the twitter monitor work in three step, the first two step it detects the trend and in the third step, it performed the analysis. In the first step they find the burst keywords (suddenly appear) after that they extract other information's from tweets belong to the trend and try to find the interesting relation.

The Working of the twitter monitor can be explained by the help of flow diagram given below.

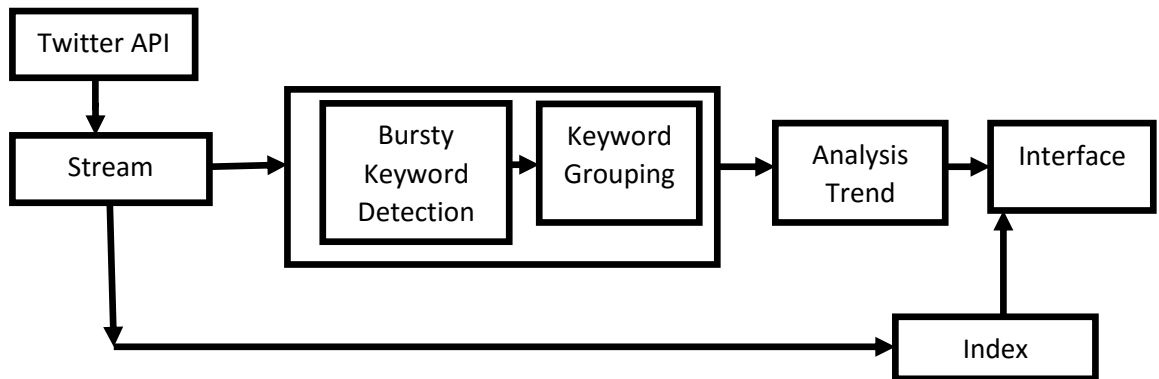


Figure 2.1: Twitter Monitor Architecture

The algorithm used by the twitter monitor for the detection of burst keywords is The one-pass algorithm, Queue Burst.

**One-Pass algorithm:** In registering, a one-pass calculation is one which peruses its information precisely once, all together, without unbounded buffering. A one-pass calculation, for the most part, requires  $O(n)$  (see 'huge O' documentation) time and not exactly  $O(n)$  stockpiling, here  $n$  is the measure of input. Basically one-pass calculation works as takes after: the article depictions are handled serially; the primary item turns into the group illustrative of the principal bunch; each ensuing article is coordinated against all group agents existing at its preparing time; a given item is allocated to one bunch (or increasingly if cover is permitted) by condition on the coordinating capacity; when an item is relegated to a group the delegate for that group is recomputed; if an article fizzles a specific test it turns into the bunch illustrative of another group.

E.Agichtein et al. [2] explained how we get the high quality of content or data from the social media. The nature of client produced content shifts drastically from great to mishandle and spam, with the assistance of this paper the author attempt to present a general order structure for joining the proof from various wellsprings of data that changed over into online networking sort and quality definition. The primary way is group entryway in which clients answer the question asked by some other client which gives an option channel to getting data on the web, as opposed to looking in the program. The primary way is group entryway in which clients answer the question

asked by some other client which gives an option channel to getting data on the web, as opposed to looking in a program of question- answering system. For example, a quality score can be utilized as a contribution to positioning calculations. On an abnormal state, their approach is to endeavor elements of online networking that are naturally associated with quality, and after that prepare a classifier to suitably choosing and weight the components for every particular kind of thing, errand, and quality definition. A quality score can be utilized as a contribution to positioning calculations. On an abnormal state, their approach is to endeavor components of online networking that are naturally connected with quality, and afterward prepare a classifier to properly choose and weight the elements for every particular sort of thing, errand, and quality definition. All these feature types are used as an input to a classifier that can be tuned for the quality definition for the particular media type. In the next section, we will expand and refine the feature set specifically to match our main application domain of community question/answering portals.

F.benevenuto et al. [5] explain how we get the recent hot topics from the source of huge amount of real time data i.e. from Twitter. The earlier topic mining from Internet Web pages are based on text clustering, however, the comparison with web content and the twitter message is short and time varying which is a great challenge for a flexible stream mining of hot topics. This paper proposes a Pattern mining algorithm (i.e. FP- Stream) to detect the latest topic from twitter stream. The key technology used by this paper is classification, clustering and tracking, sentiment tendency identification and multi-documentation summarization. The major modules of the paper are explaining as follows: -

- (i) **Data Acquisition:** -This module plays an important and major role in the system, this module downloads web pages from the internet after that they feed into the web text preprocess module for data cleaning.
- (ii) **Web Text Preprocess:** -This module collects the data which was downloaded from the internet and cleans the data i.e. remove the noise and after that, the data can be passed into next module.
- (iii) **Text Vectorization Module:** -This module also plays a major role in the system, the clean data which was received from web text preprocess module was converted into numeric vectors and pass to text classification module.

- (iv) **Text Classification:** -After that, every one of the information will be ordered into predefined classes in the content bunching module where every class speaks to a theme, this module utilizes distinctive calculation will yield diverse subjects.
- (v) **Hot Topic Evaluation:** -This module will give positions to the distinctive subject which were met by the above arrangement module by thoroughly concentrate different parameter, for example, report number, remark number, click number and so on. Finally, step the intriguing issue module produces the hotly debated issue from various view focuses

**FP- Stream Algorithm:** -The FP-stream calculation is equipped for keeping up time-touchy continuous examples in information stream with the earth of restricted principle memory, this calculation just works with the lump of information i.e. it works just when enough approaching exchange have landed to shape another lump of information. The overall summary of this paper is that the latest topic detection from twitter is a challenging research work in the stream mining community and the main algorithm that is used in the stream mining was clustering, classification, FP-Stream, and K-mean.

R.Jin et al. [12] by the help of this paper the author tries to provide an improved K-mean algorithm for the data clustering. Clustering is a technology that can consider as the most important unsupervised learning problem; it is used to finding a structure in a collection of an unlabeled data. Therefore, a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. K-mean is a very simple technique which is used to measure the intra-cluster distance, the main disadvantage of K-Mean algorithm is that it can never tell the number of clusters required for data mining by the help of this paper the author try to solve this short come of K-mean algorithm and proposed a modified algorithm which provides the number of clusters required for the mining.

Bin Wen et al. [8] proposed the work of sentimental orientation classification of text and words and then try to develop a strategy for figuring word's opinion esteem has been enhanced the premise of existed semantic likenesses of HowNet Knowledge database and afterward a circuit of transductive Learning and progressed wistful introduction processing be talked about for looking better outcome. They made two-word bunches named as positive and negative and put the words taken from the

information set(tweets) into these gatherings which depend on the wire transductive learning with semantic perception.

L.wikarsa et al. [9] proposed three main phases of text mining i.e. preprocessing, processing and validation. They divide their whole work into different phases, in the first phase they collect data from the tweeter and the preprocess it to design classifier apply the naïve Bayes classifier model to create different classifiers  $P(c|d)$   $P(c)\pi P(tk|c)mk = 1$ . In brief the design of the model can be explained as follows firstly we collect the tweets from the tweeter API and after that we perform preprocessing which involves different process such as folding, cleaning, stop word removal, emotion conversion tokenization etc. when the preprocessing is done we move to use training dataset to generate model base which is based on Naïve Bayes algorithms which produce a better result with accuracy.

$$V = \max(V_j | a_1, a_2, \dots, a_n).$$

$$V = \max P(a_1, a_2, \dots, a_n | V_j) / P(a_1, a_2, \dots, a_n).$$

P.Tripathi et al. [7] explain the whole process is same up to a collection of data and preprocessing and after that they use different technique i.e. they use naïve Bayes algorithm and K-Nearest Neighbors algorithms for the sentimental analysis. Naïve Bayes theorem is used to make the classifier because it can handle an arbitrary number of independent variables whether continuous or categorical whereas the K-Nearest Neighbors method is used to make the prediction based on the outcome of the K-Nearest Neighbor algorithms to define a matrix.  $D(x, p) = \sqrt{(x - p)^2}$ , this equation is known as the Euclidean distance which is used for finding the path between the cluster/classifier. Both the classifier predicts the label of testing of the dataset and the result was compared the result of both the label and the author found that the analysis done by the use of K-Nearest Neighbor is provided more accurate and fast result as compared to other.

R.Batool et al. [10] proposed a new methodology for sentimental examination by applying the information enhancer and equivalent word fastener module which expanded the pickup range to 0.1% to 55%. The extracted data was stored in ontologies like SIOC, FOAF, and OPO etc. They use twitter API to fetch the data from twitter in the XML format which is proposed to remove the stop words, then the

proposed tweets given to the Alchemy API process is by natural language processing and machine learning technique to get the keywords and user sentiments this task was done under the module of knowledge enhancer, in the second part of the proposed system was synonym binder in which the system binds the synonyms with each entity and extract the keyword from the Knowledge enhancer. In this paper, the author demonstrated the system which extracts data from the tweets and gives sentiments of the user and the whole task is one three different phase.

X. Wang et al. [17] introduce the concept to identify the user's sentimental features by the help of some parameters such as affection, sentiments, and user attributes again these attributes are divided into a subclass to make the different sets. The creator utilized the four fundamental law to finish up the aftereffect of the examination as the general obliges of human suppositions in view of notion space, the law utilized by the creator as a part of this paper are Law of nostalgic dormancy, law of starting point asymptotically steady, law of wistful confliction and law of wistful dissemination. Along these lines, the creator can search out the identities of each web clients, for example, controllability and transmissibility, and finally, the case-based review is done in this paper to show the aftereffect of the investigation.

P.K Singh et al. [16] explained how the mining and visualization of social media data can be helpful in the making of the market decision, they present a method that aims to understand the need of the market for a company and try to develop the system/tool which can support it, the work was divided into two use case with Swiss company, the whole methodology used and defined in this paper was based on Cross-industry standard process of data mining. The development process contains seven-step and the work start with the identification of client needs through interview the remaining steps are identification of goal, define of precondition, define of post condition, describe of main flow and the last step is description of exception and the help of this they created a tool which provides the analysis result based on the user feedbacks and comment which help to make decision/market planning to improve the business process

Tina R.Patil et al. [31] explain the analysis of the performance evaluation based on the correct and incorrect instance of data classification using Naïve Bayes and J48 tree, in which the Naïve Bayes is based on the probability concept whereas the J48 is based

on the decision tree. In this paper, the comparison is shown in the experimental result based on the following parameter i.e. accuracy, true positive rate, sensitivity, and specificity. In this, they represent that the correct instance generated by J48 is much greater than the Naïve Bayes which means that J48 is the simple decision tree classification which provides the good result as compared to Naïve Bayes but in term of cost analysis, both of the method i.e. Naïve Bayes and J48 has the same output.

Young Gyo Jung et al. [25] explain the concept of enhancing naïve Bayes classifier for real time sentimental analysis using spark R. In this paper they proposed the concept of Laplace smoothing technique with binaries NBC for improving the accuracy, and engaging Spark R for speed-up via distributed and parallel processing. The author divides the whole methodology of their work into two sections. The first section describes the spark R, Naïve Bayes, monomial Naïve Bayes and their relative study whereas in the next section they describe the work for improving the existing schemes and implemented in the spark tool. In their proposed system they focus on the Laplace smoothing for the fast and better result, the Naïve Bayes is totally worked on the probability concept which may decrease the accuracy so by applying the concept of Laplace smoothing the accuracy of the sentimental analysis will increase.

Shiju Sathyadevan et al. [22] discuss a method which is based on the concept of Naïve Bayes algorithm to categorize the given document pressed into “Cloud”. They implemented the algorithm by calculating the weight which gives a better and fast result. By their experimental result, they show that the result which gets by using the concept of converged weight is much better than the previous one, for the result comparison they used accuracy, recall value, precision value and the time taken by the system to classify the system. They also suggest that this concept will be applied with NOSQL for handling the bulk amount of data and for fast execution.

Shweta Rana et al. [23] discuss the comparative analysis of the algorithm of Naïve Bayes and support vector machine (SVM) which is generally used in the sentimental analysis. In their study, they collect the user review from a different source and apply the both of the algorithms i.e. Naïve Bayes and SVM on the same data set and on the basis of result they found that in their study the support vector machine is better than the Naïve Bayes in term of accuracy and speed. For the comparison of the result, they take different parameters likes accuracy, recall or true positive rate value and

precision value and on this basis, they conclude that SVM is better than the Naïve Bayes for their data set.

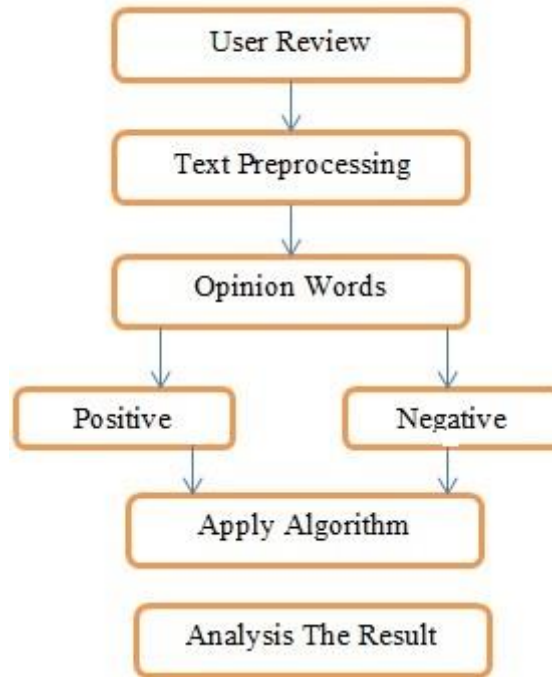


Figure 2.2: System Used

Prashast Kumar et al. [20] explain a method of feature precise opinion mining and sentimental analysis across e-commerce website, the main aim of their study is to automate the process of collecting the reviews of any product given by the end user apply the algorithm to analysis the opinion and the sentiments of the user for that product individually. Their work involves the filtering the unusual review and then provide the qualification of the thousand review of a particular product analysis this and finally provide a summarized data to the end user which will help for making the correct decision making.

Vivek Sharma et al. [21] discuss the sentimental mining and classification of music lyrics using SentiWordNet, the main aim of the author is to analysis the sentiments of the lyrics of the song and classified that whether the songs is suitable for the audience or not by labeling positive and negative. For their research they used the SentiWordNet which provide a platform for the analysis of text, in their proposed word they divide the algorithm into two parts, first part of the algorithm explain the



preprocessing of the data and in the second part of the algorithm is to analysis the sentiments of the lyrics and provide a label of positive or negative. In their research work, they take the textual part of the lyrics for analysis that whether the song is suitable for children, young or not.

Pankaj Kumar et al. [24] proposed a concept for the enterprise analysis through the opinion mining, in their work they collect the tweets from the microblogging site (twitter) pre-process it by applying different algorithm and process and decide the attitude of the mass in term of positive, negative or neutral towards the enterprise and the product provided by them. They use the R tool for their analysis work and apply the Naïve Bayes algorithm for the better result since R provide the better environment and tools for analysis task.

Ulrik Franke et al. [26] proposed a prospect of building a system that helps the user to detect the deception on twitter. The commitment has a few phases: First, utilizing a formerly distributed scientific categorization of digital double dealing, we inspected how Twitter can be utilized as a vector for misleading. In spite of the fact that the cases distinguished are presumably not thorough, it is sensible to trust that vast and pertinent segments of potential double dealing methodologies were found. Second, we presented the wide thought of pointers. These are the hint that user analysis the characteristic of deception. In the third phase, they discuss the indicator as the background the prospect of detecting is discussed.

Menara P Anto et al. [27] proposed a system which provides the rating to a particular product based on the sentimental analysis. Feedback helps the producer to improve their quality but many times the user does not provide the feedback related to any product will create a difficult for the manufacturer to resolve this problem the author try to develop a system which provides an automatic feedback to the product on the basis of data collected from the twitter. The data which is collected from the twitter is a filter, analysis and then generated the feedback In their implementation the author used support vector machine for making the classification of data as well as the concept of dual prediction, unigram approach etc.

F.Santos-Sanchez et al. [32] were proposed a novel framework to get information of interest from a web crawler by examining the loving and wistful substance of each

page. This data can give another approach to recognizing the group reaction to an item and given the positive versus negative normal sentimental rates in the discussions, it is conceivable to know the Web people group recognition to the item. This arrangement makes possible to recognize the best deals alternatives given to a specific item, and it additionally allows getting an incredible decrease in the required time to locate the best choices by gathering the examination pages for a simple surfing. They trust that the novel procedure can be the base in which a heavy portion of these frameworks can be customized to supply answers for the clients in the need of techniques of opinion analysis.

Masahiro Ohmura et al. [30] explain the concept of the mood extraction of a person by analyzing the thoughts or information updated by him/her on the social media site like twitter. By managing tweets posted in a large portion of a day as an information record, the technique utilizes Latent Dirichlet Allocation (LDA) to concentrate social opinions, some of which match with our everyday feelings. The detached theories, in any case, demonstrate carried affectability down to changes in time, which proposes that they are not appropriate for predicting everyday social or financial occasions. Utilizing LDA for the delegate 72 descriptive words to which each of the 800 modifiers maps while protecting word frequencies licenses us to get social conclusions that show enhanced affect ability to changes in time.

Salma Jamoussi et al. [29] proposed a concept of dynamic construction of dictionary for the sentiment classification, in their work they explain a program strategy to make the positive and negative word references that accomplishes the feelings images (emoticons, acronyms and shout words) display in remarks. All the more significantly, their thought permits to increase these word references with an enhancement step. At long last, by utilizing these readied word references, we anticipate the positive and negative polarities of the remark.

Takeru Yokoi et al. [28] proposed a method of extraction of emotion based on the eye character and symmetric string. For their work they used two approaches, the first approach is the use of natural language processing (NLP) and the second approach is the use of symmetric of the emotion of string. To extract the emotion they use to segment or component, the first segment deals with an eye representation incorporated into a lot of emotions, and in the second segment, they look for the string

symmetry of its outside and choice of the emoticon run in view of the second-arrange polynomial. The proposed approach has normal to accurately extract over 80% eye depictions that are viewed as the emotion center. Be that as it may, it has quite recently prevailing to accurately remove around 60% emoticons.

## CHAPTER 3

### PRESENT WORK

---

In this section, we split this into three subsections. The first segment i.e. 3.1 explains the problem formulation, in the next section i.e. 3.2, explain the objective of the research and in the last, and we explain the methodology of our research work.

#### 3.1 PROBLEM FORMULATION

In our research, we focus on the text mining which comes under the concept of data mining field. The term text mining is based on the developing of facts from an unstructured database or from the word-based database.

The sentimental analysis on social media has a good scope at present as well in future. It is a process of finding/determining a hidden emotional tone behind the set of words which is used to gain an understanding of the opinion, emotions, behavior of the people expressed in social media. An application of opinion analysis is very broad in a current day it is used everywhere such as in business, science, politics, social etc. Sentimental analysis has been more than only a social logical instrument. It's been an intriguing field of study. The sentimental analysis involves the direct or indirect conversation between users and provides a feedback on that basis the decision should be taken by any organization which increases the customer as well as the trust. The calculations have not possessed the capacity to anticipate with more than 60% precision the sentiments depicted by individuals. So the analysis of social media data to predict the future or any uncertainty it is a great area of research in now a day.

The primary motive of our research work is to propose an algorithm to optimize and increase the accuracy of the sentimental analysis. The algorithm is motivated from the concepts of local dependence distribution i.e. a probability based algorithm which is based on Naïve Bayes classification.

“Naive Bayes is a basic system for building classifiers: models that appoint class marks to issue occurrences spoke to as vectors of highlight qualities, where the class names are drawn from some limited set. It is not a self-contained calculation for

preparing such classifiers, but rather a group of calculations in view of a typical rule: all naïve Bayes classifiers accept that the estimation of a specific element is free of the estimation of whatever other component, given the class variable” In our work, we have suggested the new algorithm for the enhancing the present Naïve Bayes classification in which we introduced the concept of the joint probability distribution for better performance and increasing the accuracy rate of the classifier.

### **3.2 OBJECTIVE**

After considering the whole scenario, our main aim is to develop a sentimental analysis system which will handle the large social media data and then check the mood and opinion of the user and predict the condition. Our research is focused on following objectives:

- To study and analyses the existing algorithm on sentiment analysis on social media data.
- To develop the enhance Naïve Bayes classification and implement it using the twitter based data for classifying the mood and opinion of the users into certain categories.

### **3.3 METHODOLOGY**

In the approach segment, we have discussed the research methodology, development tools for our work and the flow chart.

#### **3.3.1 RESEARCH METHODOLOGY**

The primary motive of the proposed algorithm is to optimize and increase the accuracy of the sentimental analysis. The algorithm is motivated from the concepts of local dependence distribution i.e. a probability based algorithm which is based on Naïve Bayes classification.

“Naive Bayes is a basic system for building classifiers: models that appoint class marks to issue occurrences spoke to as vectors of highlight qualities, where the class names are drawn from some limited set. It is not a self-contained calculation for preparing such classifiers, but rather a group of calculations in view of a typical rule: all naïve Bayes classifiers accept that the estimation of a specific element is free of the estimation of whatever another component, given the class variable”

The equation is:

$$P(\text{Word}|\text{Count}) = \text{count}(\text{Word}, \text{Count}) + 1 / \text{count}(\text{C}) + V \text{ -----(i)}$$

$$Pa(C|E) = Pa(E|C) Pa(C) / Pa(E)$$

Where  $E = (x_1, x_2, x_3, \dots, x_n)$

$$F(E) = P_b(C = +|E) / P_b(C = -|E) \geq 1 \text{ -----(ii)}$$

The performance of the system which is used for the analysis of the sentiments of the user depends on two factors. These two factors are:

- Compatibility factor of a stream data with the memory of the system and the data structure and algorithm is used.
- Data locality and the dependency of the dataset among them self

The principle of this algorithm is

For a node  $X$ , the local dependence derivative of  $X$  in classes  $+$  and  $-$  are defined as below.

$$dd_G^+(x|p_b(x)) = p(x|p_b(x), +) / p(x|+)$$

$$dd_G^-(x|p_b(x)) = p(x|p_b(x), -) / p(x|-)$$

Basically,  $dd_G^+(x|p_b(x))$  mirrors the quality of the neighborhood of hub  $X$  in class  $+$ , which measures the impact of  $X$ 's nearby local dependency on the characterization in class  $+$ .  $dd_G^-(x|p_b(x))$  is similar. Further we have the following results.

1. When  $X$  has no parent, then

$$dd_G^+(x|p_b(x)) = dd_G^-(x|p_b(x)) = 1.$$

2. When  $dd_G^+(x|p_b(x)) \geq 1$ ,  $X$ 's local dependence in class  $+$  supports the classification of  $C = +$ . Otherwise, the classification of  $C = -$ .

Similarly, when  $dd_G^-(x|p_b(x)) \geq 1$ , local dependence in class negative supports the classification of  $C = -$ . Otherwise, it supports the classification of  $C = +$ .

The Conditional autonomy supposition is once in a while valid in the truest application, so if we link all the attributes with each other and apply joint probability we get the effective result which can be represented by the help of this equation:

$$P(X_1, \dots, X_n, C) = P(C) \prod_{i=1}^n P(X_i | Pa(X_i), C) \text{ Where } P(X_i) = \text{Parent of } X_i$$

Now if we apply log over equation (ii) we get the better result as compared to Naïve BayesLog  $f(E) = \text{Log } P(C = +|E) / P(C = -|E) \geq 0 \text{ -----(iii)}$

$f(x_1, x_2, \dots, x_n) = f_{nb}(x_1, x_2, \dots, x_n) \prod_{i=1}^n ddrG(x_i)$ , where  $\prod_{i=1}^n ddrG(x_i)$  is called the dependence distribution factor.

The pseudo code of the work is as follows:

Procedure Z (V,C,S,Qv)

Inputs

V: variables

C: factors show conditional probabilities

S: observations of values

Qv: query

Output

Posterior distribution on Qv

Local

A: take a collection of factor.

$A \leftarrow C$

for each  $Y \in V - \{Qv\}$  using elimination ordering do

if (Y is observed) then

for each  $E \in A$  involves Y do

set Y in E to observed value in O

Project E onto remaining variables

Else

$R \leftarrow \{E \in A: E \text{ involves } Y\}$

Let T be the product of the factors in R

$N \leftarrow \sum Y J$

$A \leftarrow A \setminus R \cup \{N\}$

suppose J be the product of the factors in A

$N \leftarrow \sum Q J$

Return J/N

By this way, we perform our task to improve the performance of the result of Naïve Bayes algorithms.

### 3.3.2 DEVELOPMENT TOOLS (R ENVIRONMENT)

The algorithm of enhancing Naïve Bayes for the sentimental analysis is developed in the r programming language. The platform we used is the R environment which was developed by Ross Ihaka and Robert Gentleman, it is an open source programming

language and programming condition for measurable figuring and illustrations that are encouraged by the R Foundation for Statistical Computing. The R language is generally utilized among analysts and information mineworkers for creating factual programming and information investigation. Surveys, overviews of information diggers, and investigations of discerning writing databases demonstrate that R's ubiquity has expanded considerably as of late.

The abilities of R is stretched out through client made bundles, which permit specific factual methods, graphical gadgets, (for example, the ggplot2 bundle created by Hadley Wickham), import/send out capacities, detailing apparatuses (knitr, Sweave), and so forth. These bundles are created basically in R, and infrequently in Java, C, C++, and FORTRAN.

The core R execution is written in R, C, and FORTRAN, and there are several other executions aimed at improving speed or increasing extensibility.

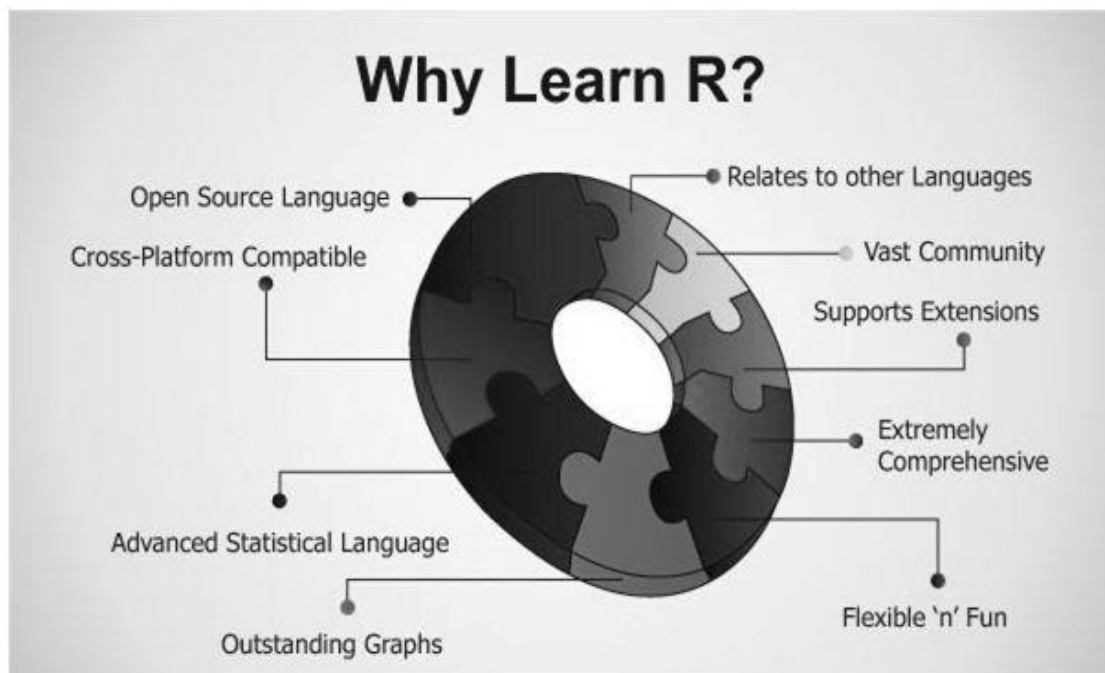


Figure 3.1: Why R for Analysis.



### 3.3.3 FLOW CHARTS OF THE WORK

The procedure is explained by the help flowchart labeled below

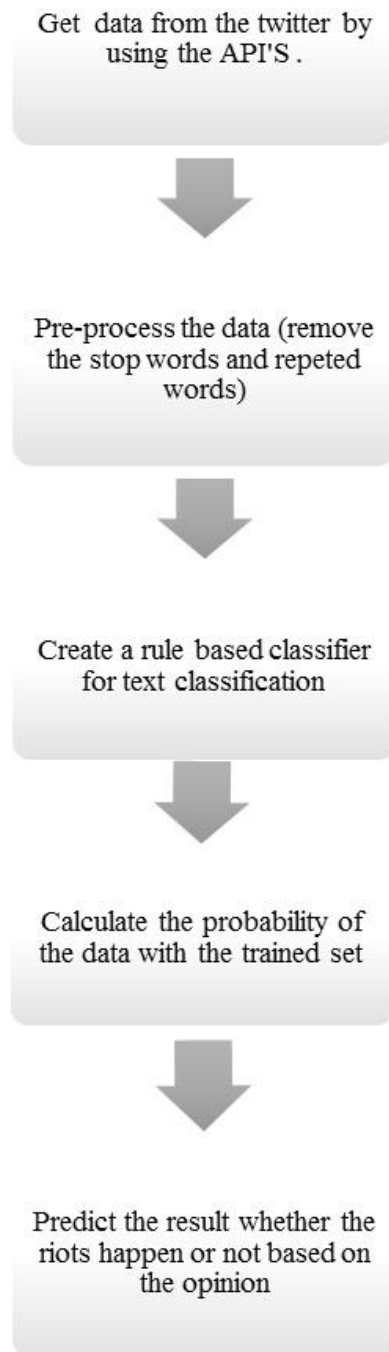


Figure 3.2 Flow Diagram of Research Methodology

#### 4.1 EXPERIMENTAL RESULT

The feasibility of any research strongly depends on the result and output obtained from the experiments. It is necessary to clear about the result at the time of formulating problems and its solution. The practicality of problem solution must be considered in the initial steps of system development. Our research is an experimental type of research in which we enhance the existing algorithm to meet our objective. There is always need practical tools for showing the hypothesis you chose in the initial phase of the research.

The expected outcome of our research work will be as follows:

- Finding the behavior of the users according to their thoughts express on social media
- Self-trained algorithm for predicting riots happens or not
- Improve the performance of analysis in term of speed and accuracy.
- Reducing the error rate by applying the concept of local dependency distribution i.e. a probability based algorithm

Therefore, general effectiveness of analysis is required to enhance through the practical executing.

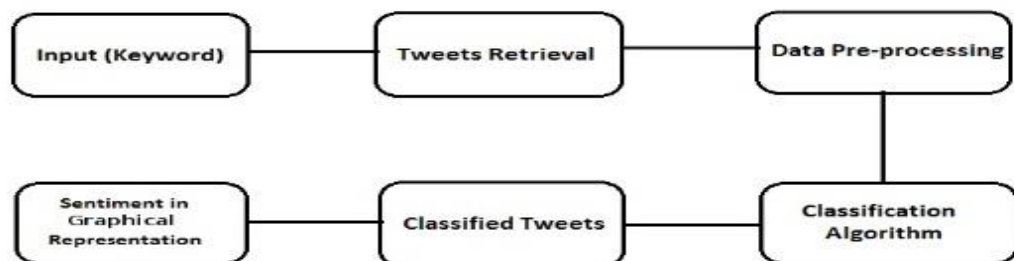
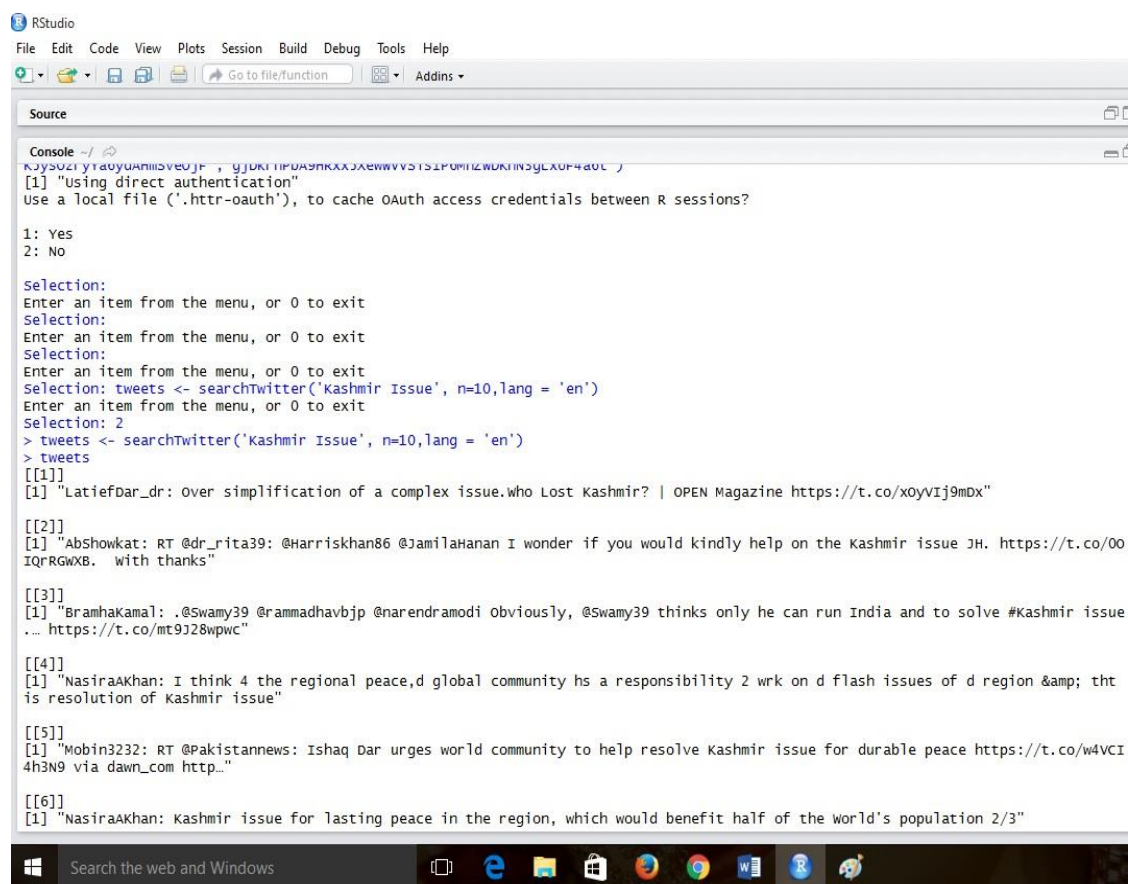


Figure 4.1: Proposed System

In our experimental work, we collect the data from the social media (twitter), preprocess it's and classified it.

To collect the data (tweets) from the twitter we used the twitter API that we created using twitter developer site and make an authentication between the R studio and the twitter site using the different packages like twitter R which make a communication channel between R and twitter and we are able to fetch the tweets for our research work.

The given figure 4.1&4.2 below will show how we take the data from the twitter for our research work.



```
RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function Addins

Source

Console ~/
R> install.packages('twitter')
[1] "Using direct authentication"
use a local file ('.httr-oauth'), to cache OAuth access credentials between R sessions?

1: Yes
2: No

Selection:
Enter an item from the menu, or 0 to exit
Selection:
Enter an item from the menu, or 0 to exit
Selection:
Enter an item from the menu, or 0 to exit
Selection: tweets <- searchTwitter('Kashmir Issue', n=10, lang = 'en')
Enter an item from the menu, or 0 to exit
Selection: 2
> tweets <- searchTwitter('Kashmir Issue', n=10, lang = 'en')
> tweets
[[1]]
[1] "LatiefDar_dr: Over simplification of a complex issue.who Lost Kashmir? | OPEN Magazine https://t.co/xoyVIj9mDx"

[[2]]
[1] "Abshowkat: RT @dr_rita39: @Harriskhan86 @Jamilahanan I wonder if you would kindly help on the Kashmir issue JH. https://t.co/00IQRGWXB. with thanks"

[[3]]
[1] "BramhaKamal: .@Swamy39 @rammadhavbjp @narendramodi obviously, @Swamy39 thinks only he can run India and to solve #Kashmir issue ... https://t.co/mt9J28wpcw"

[[4]]
[1] "NasiraAKhan: I think 4 the regional peace,d global community hs a responsibility 2 wrk on d flash issues of d region & tht is resolution of Kashmir issue"

[[5]]
[1] "Mobin3232: RT @Pakistannews: Ishaq Dar urges world community to help resolve Kashmir issue for durable peace https://t.co/w4VCI4h3N9 via dawn_com http..."

[[6]]
[1] "NasiraAKhan: Kashmir issue for lasting peace in the region, which would benefit half of the world's population 2/3"
```

Figure 4.2: Tweets fetch from twitter using R



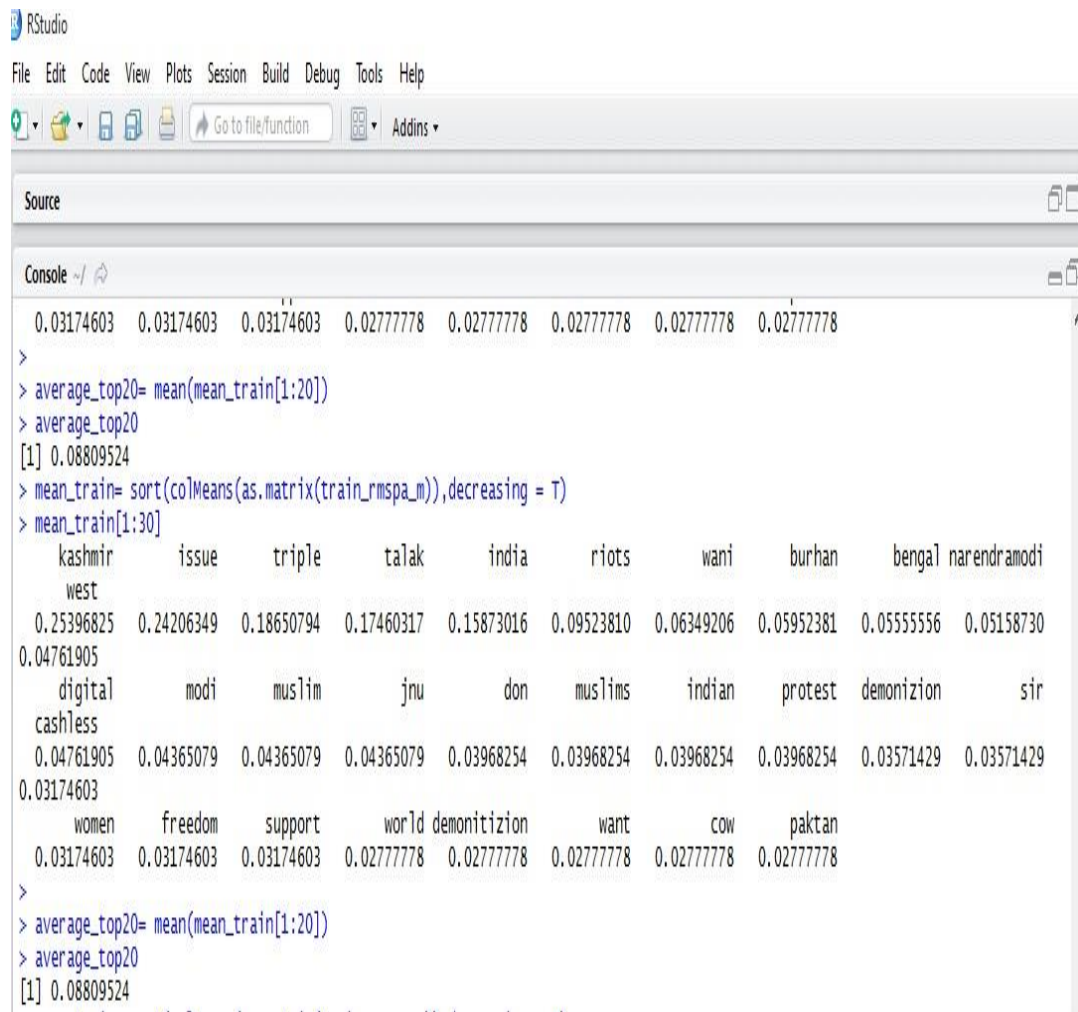


Figure 4.4: Mean of top 20 words

The Naïve Bayes algorithm is based on the probability concept, our work is also on the concept of probability so the probability of each word whether it is negative or positive is calculated by our system on that basis the prediction should be taken. The figure given below is the probability of each word whether it is negative or positive.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

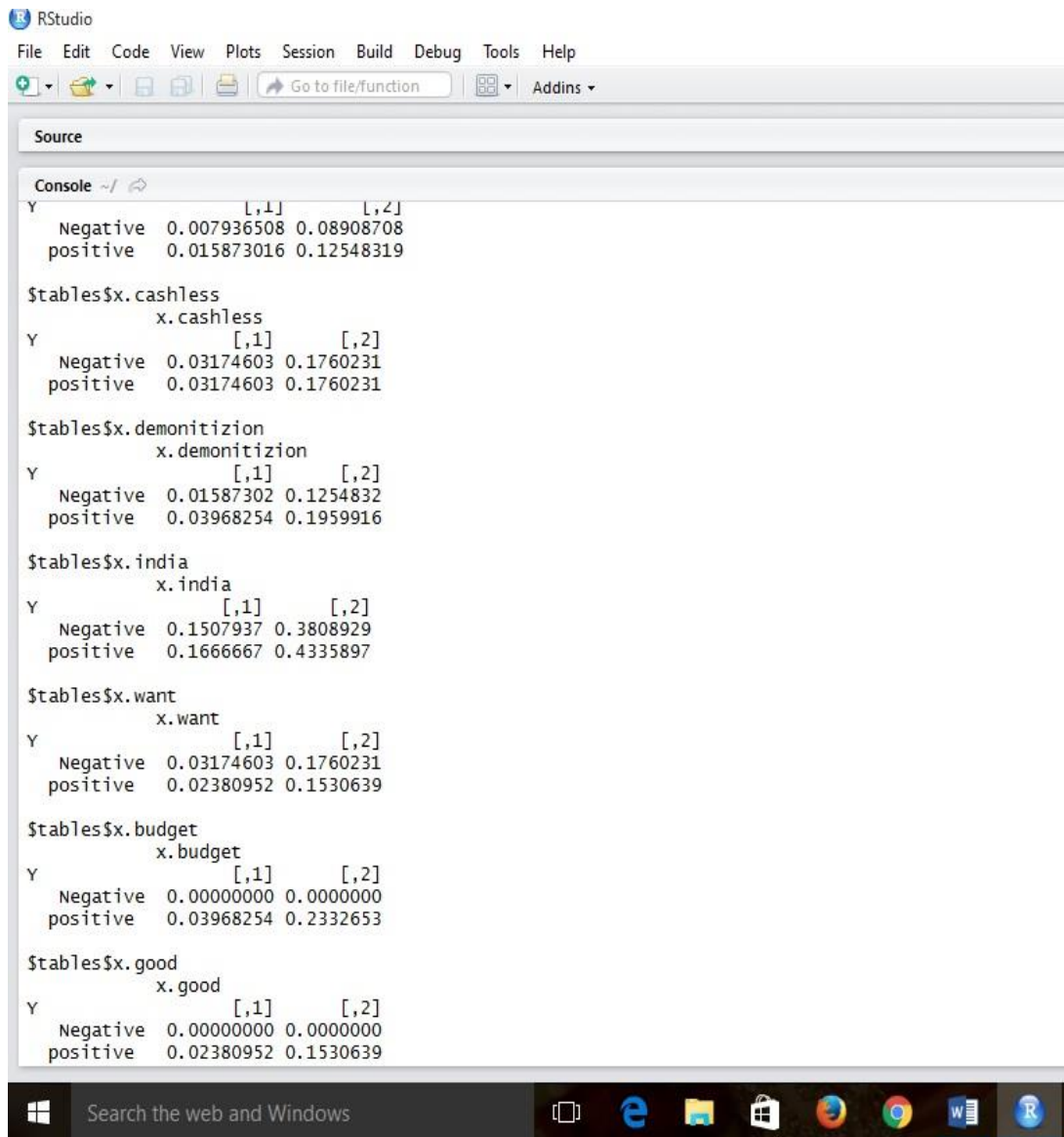


Figure 4.5: Probability of words

Word cloud is a collection of words used in the text mining in which each word has different size and colors which indicate its frequency and importance. In our work, we also represent the frequent words which are used in the data set and its frequency by the help of word cloud.

The word cloud generated by our system over the data set which we used is represented in the figure given below.

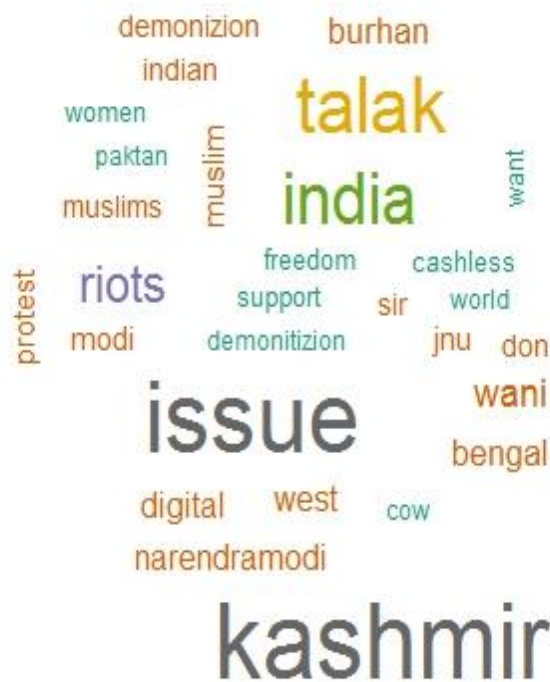


Figure 4.6: Word Cloud

Word Frequency is the representation of the most used words in the text after applying the text mining. The word frequency is the show the intensity of the words. To represent the most frequent word used in the text mining we plot a bar graph, which shows frequency of top ten words. The bar graph given in the fig.8 represents the frequency of the top 10 words which is created by our system over the given data set.

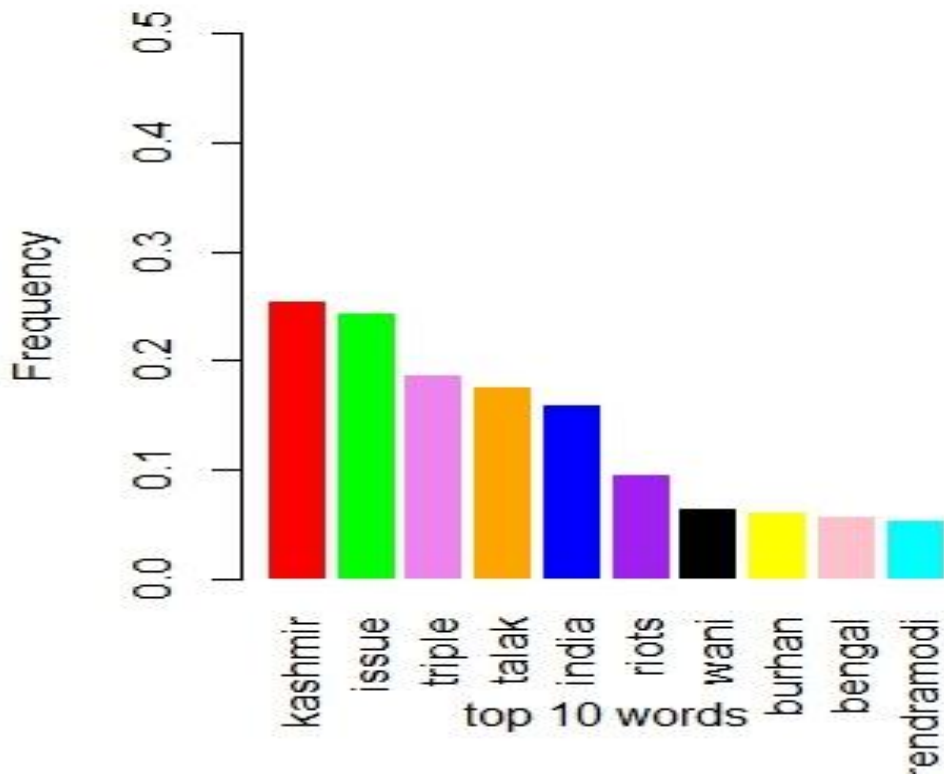


Figure 4.7: Frequency of top 10 words

## ACCURACY

The term accuracy is defined as a freedom from the error which means that your work is going in a right manner, the accuracy of the system totally depends on the process of implementation, the algorithm you used, implementing tools and the data set you taken.

As we discuss above the main aim of our work is to enhance the Naïve Bayes i.e. to increase the accuracy of the present algorithm.

The figure given below shows the accuracy of the enhance Naïve Bayes generated by our set on a given data set. As we see that the accuracy is about 68.25 which is higher than another classifier.



RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function Addins

Source

Console ~/

```
$levels
[1] " Negative " "positive"

$call
enhanceNaiveBayes.default(x = X, y = Y, laplace = laplace)

attr(,"class")
[1] "enhanceNaiveBayes"
> summary(Bow_nb_m)
      Length Class Mode
apriori    2  table  numeric
tables   150 -none- list
levels     2 -none- character
call       4 -none- call
> ptm <- proc.time()
> test1pred=(predict(Bow_nb_m,newdata= test1_data_m))
> proc.time() - ptm
      user system elapsed
 1.09   0.00   1.14
> #test1pred
> #confusionMatrix(test1pred,test1_m, positive = "Positive", dnn = c("Prediction","True"))
> mmetric(test1pred,test1_m,c("ACC", "TPR", "PRECISION", "F1"))
      ACC      TPR1      TPR2 PRECISION1 PRECISION2      F11      F12
68.253968 68.000000 100.000000 100.000000   2.439024  80.952381  4.761905
>
> library(kernlab)

Attaching package: 'kernlab'

The following object is masked from 'package:ggplot2':

    alpha

Warning message:
package 'kernlab' was built under R version 3.3.2
> Bow_svm_m= ksvm(y~.,data= train_data_m)
> Bow_svm_m
Support Vector Machine object of class "ksvm"
```

Windows Search the web and Windows

Figure 4.8: Accuracy of Enhance Naïve Bayes System

The screenshot shows the RStudio environment with the following content in the console:

```

x.jnu
      [,1]      [,2]
Negative 0.1031746 0.3539348
positive 0.0000000 0.0000000

x.protest
      [,1]      [,2]
Negative 0.08730159 0.3103505
positive 0.00000000 0.0000000

x.students
      [,1]      [,2]
Negative 0.03968254 0.2332653
positive 0.00000000 0.0000000

x.pakistan
      [,1]      [,2]
Negative 0.023809524 0.15306395
positive 0.007936508 0.08908708

> summary(Bow_nb_m)
      Length Class      Mode
apriori      2      table numeric
tables     150    -none-  list
levels       2    -none- character
call         4    -none- call

> testlpred=predict(Bow_nb_m,newdata= test1_data_m)
> mmetric(testlpred,test1_m,c("ACC", "TPR", "PRECISION", "F1"))
      ACC      TPR1      TPR2 PRECISION1 PRECISION2      F11      F12
63.492063 65.546218 28.571429 93.975904  4.651163 77.227723  8.000000
>
>
>
> |

```

Figure 4.9: Accuracy of Naïve Bayes

## 4.2 COMPARISON WITH EXISTING TECHNOLOGY

Precision and the true positive rate are the two main factors which are used to assess the classification performance. Precision is the percentage of record classified into a category that has been correctly classified. The formula used to calculate the precision value is  $TP/TP + FP$ , whereas the true positive rate or recall demonstrates the number of positive records that are correctly identified. In our work, the comparison of the precision and true positive rate of the different classifier is shown in the below.

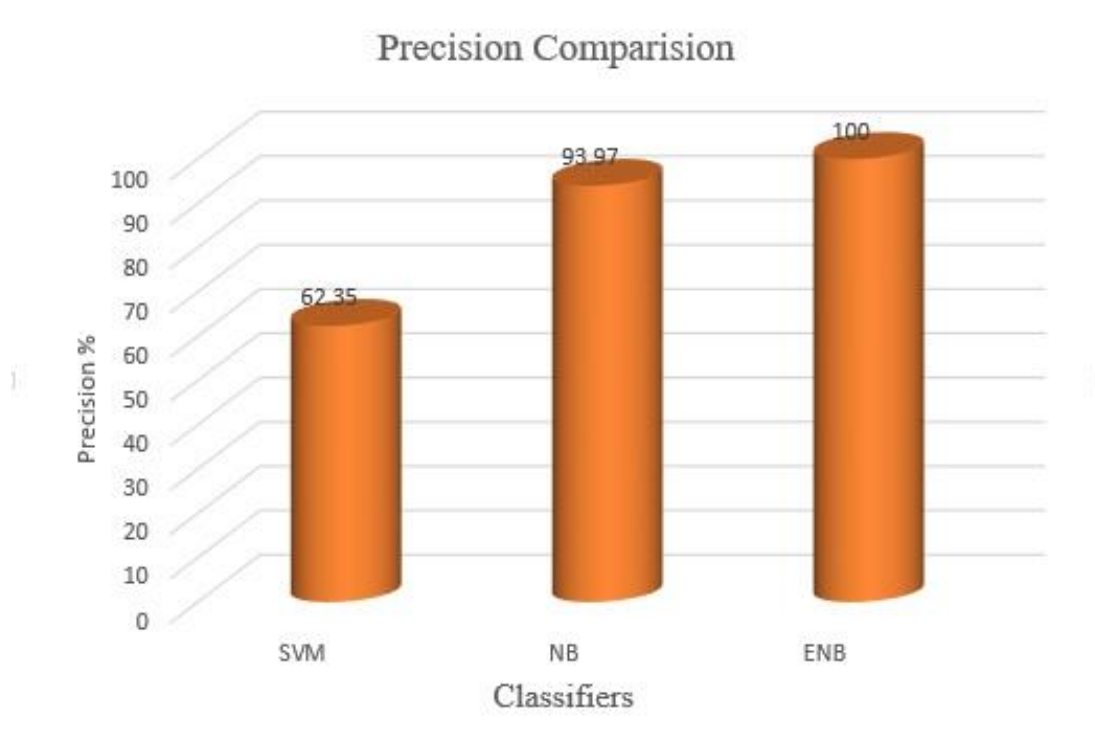


Figure 4.10: Precision Comparison

In the above figure, we compare the precision value of classifier's i.e. support vector machine, Naïve Bayes and the enhance Naïve Bayes. The result generated by the system shows that the precision value of support vector machine is 62.35, the precision value of Naïve Bayes is 93.97 whereas the precision value of enhancing Naïve Bayes is 100.

True positive rate or recall value defined as the number of positive records that are correctly identified, in another word we can say that the term true positive rate or recall is also referred as the sensitivity of the classifier.

The recall value of the different classifier like support vector machine, Naïve Bayes and enhance Naïve Bayes is represented by the help of bar graph. The true positive rate or recall value of our system can be compared to the earlier system i.e. support vector machine and Naïve Bayes and show the result in the figure given below.



Figure 4.11: TPR Comparison

The figure 4.11 shows the comparison of true positive rate or recall value generate by the system by applying the different classifiers like support vector machine (SVM), Naïve Bayes (NB) and enhance Naïve Bayes (ENB) and by the visualization we conclude that the outcome of true positive rate generated by support vector machine is and enhance Naïve Bayes is approximately same but the true positive rate of Naïve Bayes is slightly low as compared to support vector machine and enhance Naïve Bayes.

Accuracy Comparison of classifiers shows the accuracy for the individual class level. In our case, we have SVM, Naïve Bayes and Enhance Naïve Bayes are the classifier.

As we discuss above the main aim of our work is to enhance the Naïve Bayes i.e. to increase the accuracy of the present algorithm. The graph given below shows the rate of accuracy generated by the system using different classifiers i.e. support vector machine, naïve Bayes and enhance Naïve Bayes

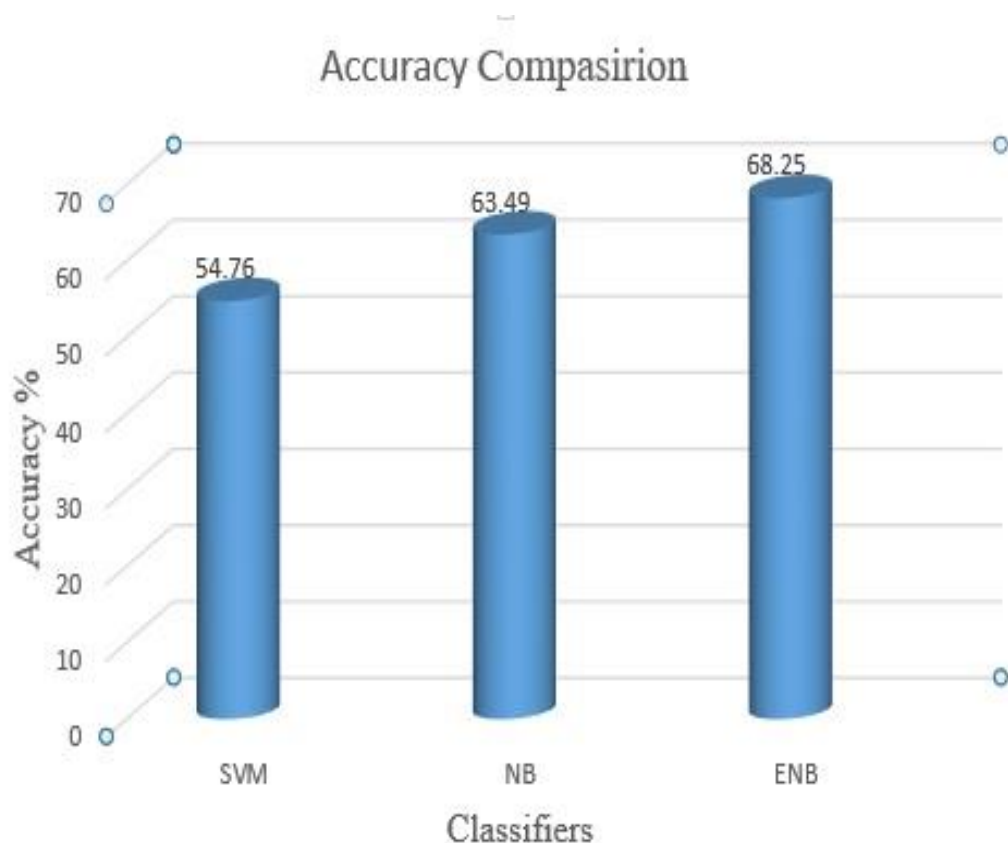


Figure 4.12: Comparison of Accuracy of Different Classifiers

The given figure 4.13 will show the comparison of the result in term of accuracy, precision value and recall value of all the classifier i.e. enhance Naïve Bayes, Naïve Bayes and Support vector machine. By the help of giving line chart below, we are able to find the accuracy, precision's value, true positive rate etc. of different classifier over the same set of data generated from the social media (twitter).

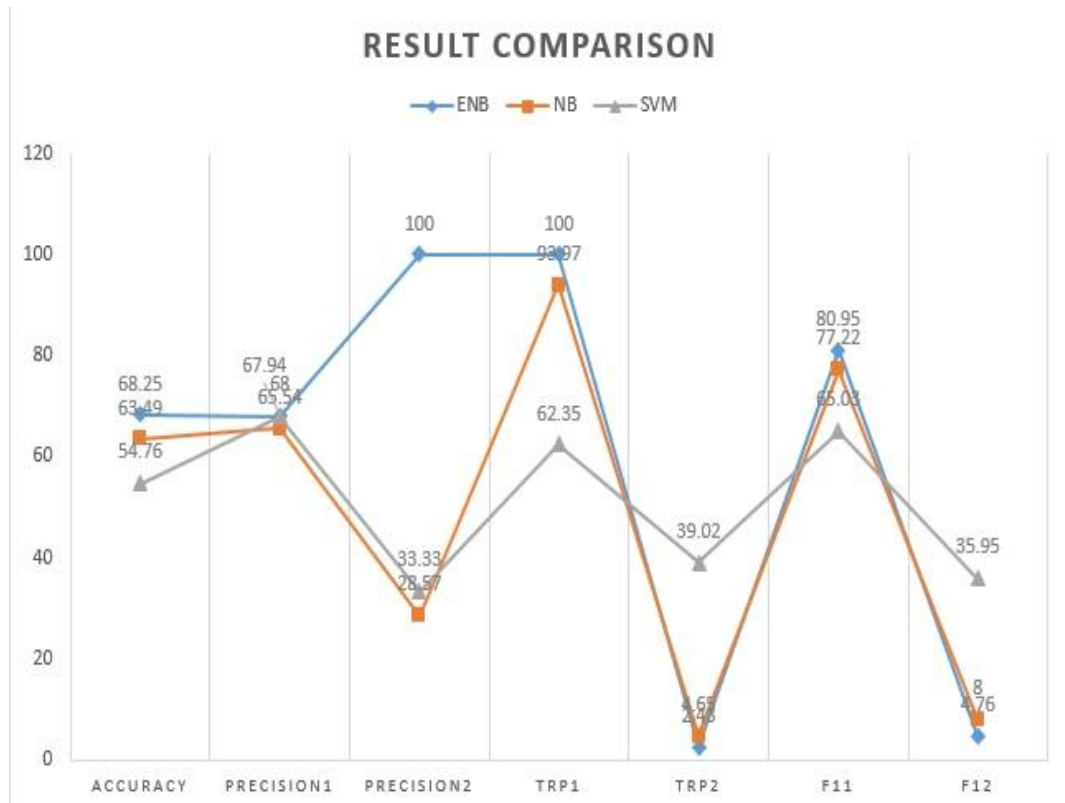


Figure 4.13: Comparative Analysis

In the above chart, the blue line represents the enhance naïve Bayes (ENB), orange line represents the Naïve Bayes (NB) whereas the gray line represents the support vector machine (SVM). The line chart shows the comparative analysis of the result generated by our system using different classifiers over the same data set, the above visualization shows that the enhance naïve Bayes algorithm is more optimized than Naïve Bayes and support vector machine in term of accuracy and true positive rate or recall vale and precision values.

## **CHAPTER 5**

### **CONCLUSION AND FUTURE SCOPE**

---

#### **5.1 CONCLUSION**

An important task of the algorithm is to analysis the sentiments of the users according to their thoughts which are generally updated by the users on social media and predict the result. From the literature review we came to realize that sentiment analysis/opinion mining especially for micro-blogging platform is still in developing stage and lots of future work can possible, so we proposed an idea of local dependency distribution over probability based algorithm which will be further improved the performance of the system. The main focus of the study is to analysis the recent news and what is the different opinion of the people, whether their thoughts will aggressive/ normal as per the situation of the country and the decision taken by government which leads or not to riots. The sentimental analysis involves the direct indirect conversation between users and provides a feedback on that basis the decision should be taken by any organization which increases the customer as well as the trust. Classification techniques will be used for keeping the track of the dataset and for making the classifiers in the system for differentiating the positive and negative words. Apart from this many other mathematical theorems also are used based on the probability for the improvement of the system.

#### **5.2 FUTURE WORK**

Our research work is to focus on the enhancing the present naïve Bayes algorithm to increase the accuracy rate of the sentimental analysis so that the outcome is more accurate and reliable.

Text mining is the theory through which we can mine the important and useful information from the unstructured database but the importance of the information directly depends on the accuracy and recall value, so our research can further applied in the field of business analytics, online campaign etc.

## REFERENCES

---

- [1] V. M. Pradhan, J. Vala, and P. Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining," *Int. J. Comput. Appl.*, vol. 133, no. 9, pp. 7–11, 2016.
- [2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding High-quality Content in Social Media," *Int. Conf. Web Search Data Min.*, pp. 183–193, 2008.
- [3] C. Greer and E. McLaughlin, "We predict a riot?: Public order policing, new media environments and the rise of the citizen journalist," *Br. J. Criminol.*, vol. 50, no. 6, pp. 1041–1059, 2010.
- [4] S. Ahmed and A. Danti, "A novel approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using web data," *Int. Conf. Trends Autom. Commun. Comput. Technol. I-TACT 2015*, pp. 0–4, 2016.
- [5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," *Collab. Electron. Message. anti-abuse spam Conf.*, vol. 6, p. 12, 2010.
- [6] V. Martin, "Predicting the french stock market using social media analysis," *8th Int. Work. Semant. Soc. Media Adapt. Pers. SMAP 2013*, pp. 3–7, 2013.
- [7] P. Tripathi, S. K. Vishwakarma, and A. Lala, "Sentiment Analysis of English Tweets Using RapidMiner," 2015.
- [8] B. Wen, S. Duan, B. Rao, and W. Dai, "Research on Word Sentimental Classification Based on Transductive Learning," *2015 8th Int. Symp. Comput. Intell. Des.*, pp. 153–156, 2015.
- [9] L. Wikarsa and S. N. Thahir, "A text mining application of emotion classifications of Twitter's users using Na??ve Bayes method," *Proceeding 2015 1st Int. Conf. Wirel. Telemat. ICWT 2015*, 2016.
- [10] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, "Precise Tweet

Classification and Sentiment Analysis,” pp. 1–6, 2013.

- [11] K. Chai, H. T. Hn, and H. L. Cheiu, “Naive-Bayes Classification Algorithm,” *Bayesian Online Classif. Text Classic. Filter.*, pp. 97–104, 2002.
- [12] R. Jin and G. Agrawal, “An algorithm for in-core frequent itemset mining on streaming data,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 210–217, 2005.
- [13] M. Mathioudakis and N. Koudas, “Twitter monitor: trend detection over the twitter stream,” *SIGMOD '10 Proc. 2010 ACM SIGMOD Int. Conf. Manag. data*, pp. 1155–1158, 2010.
- [14] T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” *Int. J. Comput. Sci. Appl. ISSN 0974-1011*, vol. 6, no. 2, pp. 256–261, 2013.
- [15] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the Poor Assumptions of Naive Bayes Text Classifiers,” *Proc. Twent. Int. Conf. Mach. Learn.*, vol. 20, no. 1973, pp. 616–623, 2003.
- [16] P. K. Singh, A. Sachdeva, D. Mahajan, N. Pande, and A. Sharma, “An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites,” *Proc. 5th Int. Conf. Conflu. 2014 Next Gener. Inf. Technol. Summit*, no. Statement 3, pp. 329–335, 2014.
- [17] X. Wang and X. Luo, “Sentimental space based analysis of user personalized sentiments,” *Proc. - 2013 9th Int. Conf. Semant. Knowl. Grids, SKG 2013*, pp. 151–156, 2013.
- [18] H. Zhang, “The Optimality of Naive Bayes Naive Bayes and Augmented Naive Bayes,” *Am. Assoc. Artif. Intell.*, 2004.
- [19] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter,” *Proc. 20th Int. Conf. World wide web - WWW '11*, p. 675, 2011.



- [20] P. Kumar, "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites," no. Statement 3, pp. 329–335, 2014.
- [21] V. Sharma, "Sentiments Mining and Classification of Music Lyrics using SentiWordNet," pp. 1–6, 2016.
- [22] S. Sathyadevan, "Improved Document Classification Through Enhanced Naive Bayes Algorithm," pp. 100–104, 2014.
- [23] S. Rana, "Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques," no. October, pp. 106–111, 2016.
- [24] P. Kumar, K. Manocha, and H. Gupta, "Enterprise analysis through opinion mining," *2016 Int. Conf. Electr. Electron. Optim. Tech.*, pp. 3318–3323, 2016.
- [25] Y. G. Jung, K. T. Kim, B. Lee, and H. Y. Youn, "Enhanced Naive Bayes Classifier for real-time sentiment analysis with SparkR," *2016 Int. Conf. Inf. Commun. Technol. Converg.*, pp. 141–146, 2016.
- [26] U. Franke and M. Rosell, "Prospects for detecting deception on twitter," *Proc. - 2014 Int. Conf. Futur. Internet Things Cloud, FiCloud 2014*, pp. 528–533, 2014.
- [27] M. Antony, N. Johny, V. James, and A. Wilson, "PRODUCT RATING USING SENTIMENT ANALYSIS," pp. 3458–3462, 2016.
- [28] T. Yokoi, M. Kobayashi, and R. Ibrahim, "Emoticon Extraction Method Based on Eye Characters and Symmetric String," *Proc. - 2015 IEEE Int. Conf. Syst. Man, Cybern. SMC 2015*, pp. 2979–2984, 2016.
- [29] S. Jamoussi and H. Ameer, "Dynamic Construction of Dictionaries for Sentiment Classification," *2013 Int. Conf. Cloud Green Comput.*, pp. 418–425, 2013.
- [30] M. Ohmura, K. Kakusho, and T. Okadome, "Social mood extraction from twitter posts with document topic model," *ICISA 2014 - 2014 5th Int. Conf. Inf. Sci. Appl.*, pp. 2–5, 2014.

- [31] Tina R.Patil and S.S.Sherekar, “Performance analysis of Naïve Bayes and J48,classification algorithm” . - *IJCSA 2013 vol 6 No 2*.
- [32] F. S. Sanchez and A. M. Vazquez, “Sentiment analysis for e-services,” *Proc. - 2014 IIAI 3rd Int. Conf. Adv. Appl. Informatics, IIAI-AAI 2014*, pp. 42–47,

# APPENDIX A

## ABBREVIATIONS

---

**viz.:** As follows

**i.e.:** That is

**CRISP-DM** Cross-industry standard procedure of information mining

**CFO:** Central Force Optimization

**FP:** Frequent pattern

**SVM:** Support Vector Machine

**NV:** Naïve Bayes

**D- Tree:** Decision Tree

**SNA:** Social Networking Analysis

**NB:** Naïve Bayes

**ENB:** Enhance Naïve Bayes

**TRP:** True Positive rate.

**TP:** True Positive

**FP:** False Positive

**NLP:** Natural Language Processing.