A HybridKNN-Naïve Baye's Algorithm using TF-IDF to Sentiment Analysis on Twitter Data

Shah Shaheena Ashraf[1], Sahil Verma[2], Kavita[3]

[1]Student, M.Tech, EMGOI, Ambhala
[2]Assitant Professor, Dept. Of CSE, E-Max Group of Institutes, Ambhala
[3]Assitant Professor, Dept. Of CSE, E-Max Group of Institutes, Ambhala

***Abstract*: Social media has attracted the attention of researchers across the globe in current times. The reason can be attributed to the large set of data available due to active involvement of the users on such platforms. The thesis proposes a novel strategy of sentiment analysis on twitter data using hybrid algorithm. Analysis of public mood regarding a specific topic is a complex task which involves many aspects like preprocessing, score calculation, classification algorithm etc. The thesis proposes a novel strategy where the effect of other tweets for score calculation are taken into account. Also the grammatical mistakes and the location influence of tweet origin are taken into account for pre-processing. A hybrid KNN-Naïve Baye's algorithm has been developed which will address the short comings of earlier used algorithms for handling high dimensionality data through chi-square technique. The results will be equated to that of traditional algorithms regarding precision, accuracy and recall values.**

***Keywords*: Sentiment Analysis, KNN, Hybrid Algorithm, Feature Selection**

1. Introduction

Social media is a great medium for exploring developments which matter most to a broad audience and it is the platform where people exchange, share and create information and ideas in virtual groups and networks. Social media platforms and technologies exist in many different forms which include weblogs, micro blogging, magazines, Internet forums, social blogs, social network, wiki, podcasts, pictures or photographs, Video, rating and social bookmarking. Micro blogging websites have evolved to become source of varied kind of information. Micro blogs have become a trending platform where people post real time posts which reveal their opinions on a range of topics, discuss recent problems, complain, and express positive sentiment for products they use in everyday life. In fact, companies manufacturing such products have started to poll these micro blogs to get a sense of general sentiment for their product. Companies analyze user's messages and reactions and give a suitable feedback to users on microblogs. Social media continuously advance in popularity as well as in importance in modern society. Different type of opinions including both positive and negative, public and private about a variedrange of themes are communicated and spread progressively via several social media platforms, with twitter being amongst the timeliest. Social media has become unique biggest forums to express ones opinion. Sentiment analysis can be defined as the text classification method which is used to extract information existinginside the text. This extracted information can be then additional categorized on the basis of its polarity as positive, negative or neutral. It can also be termed as a computational task of mining sentiments from the opinion. Some opinions represent sentiments and some opinions do not represent any sentiment.

1.1 Sentiments

Social media sites such as Twitter, Face book, and You-Tube offers great platform for students to share their experiences, vent emotion and stress, and pursue social provision on several social media sites; students debate, discuss and express their day to day happenstances in a casual and informal manner. Students' digital footprints has offered wide range of implicit knowledge and a completely novel viewpoint for educational scholars, researchers and specialists to understand students' experiences outside the controlled classroom environment. This understanding can inform institutional decision-making on interventions for at-risk students, improvement of education quality, and thus enhance student recruitment, retention, and success.

Twitter is a most prevalent trending social media platform. Its content is open to all public and very brief (up to 140 characters per tweet). Twitter offers free APIs which are used for streaming the data. Therefore, I have started from analyzing students' posts on Twitter.APIs is developed to access Twitter data that can be divided into two types in terms of their design and access technique: REST APIs are grounded on the REST architecture2 currently used for designing web APIs. Data retrieval is done by these APIs using pull strategy. To assemble information a user must explicitly request it. Streaming APIs offers a continuous, constant stream of public information extracted from Twitter. For data retrieving push strategy is used by these APIs. After a request for information is made, the Streaming APIs deliver a continuous brook of updates with no further input from the user.

The goal of sentiment analysis is to define the approach of a user (writer/speaker) in accordance to a specific subject or the total contextual polarity of a certain document. The approach may be his or her evaluation or judgment, affective state (the emotive situation of the user when writing), or the proposed emotional communication (the emotional effect) the author desires to experience the same effect on the reader. An elementary task in sentiment analysis is organizing the polarity of a certain text at the document, sentence, or feature/aspect level—the expressed Mining job whose job is to achieve and extract writers feelings conveyed in positive or negative comments, questions by analyzing a bulky amount of documents. An opinion is a quadruple (g, h, s, t), where g is the opinion (or sentiment) target, h is the opinion holder, s is the sentiment about the target, and t is the time when the opinion was expressed. Sentiment analysis is a current area of research in text mining field. Sentiment analysis is the computational revision of Opinions, sentiments, subjectivity toward an entity. The entity can signify individuals, actions, events or a specific topic. The two expressions sentiment analysis and opinion mining are interchangeable.They present a mutual meaning. But also in some contexts they have different meaning. Opinion mining mines, extracts and examines individual’s opinion about an event or an entity while sentiment analysis recognizes the sentiment expressed in a document then analyzes it. Thus, the main goal of sentiment analysis is to discoveryand find opinions, extract the sentiments they convey, and then sort their polarity. Sentiment analysis can be treated as a classification process.

2. Related Work

V. N. Khuc et al. [1] Proposed distributed system which is utilized for real time sentiment analysis. This proposed system comprises of two elements: a lexicon builder and a sentiment classifier and these elements have ability of running distributed system.Proposedapproachhasbeen executed
Executed by utilizing map reduce framework and database model. The result analysis indicates that lexicon element has enhanced quality and accuracy of the sentiment classifier.

Horakova and Marketa[2] presented a model which gathers tweets and posts from social networking sites and thus provide a view of business intelligence. In our context, two layers in the sentiment analysis tool are found, the data processing layer and sentiment analysis layer. Data processing layer is concerned with data gathering and data extraction, while sentiment analysis layer use an application to present the result of data mining.

Barnes [3] initiated to use the term systematically to present patterns and outlines of ties, includingtraditional concepts. Afterwards, many scholars extended the practice of systematic social network analysis. With the increasing development of online social networking site, online social networking analysis has become hot research topic recently.

X.chen, M.Vorvoreanu, K.Madhavan [4] suggested how social media sites data is helpful in understanding student learning experience. They collected data about student’s problems from twitter. A workflow was also established by them to integrate both qualitative analysis and large-scale data mining techniques. Their main focus was on engineering student’s twitter posts to understand issues and problems in their educational experiences [1]. They used Naive Bayes Multi-Label Classifiers for tweets classification then after that they compare the result of Naïve Bayes Multi-Label Classifiers with the most used and accurate classifier used in many machine learning tasks opinion in a document, a sentence or an entity feature/aspect can be positive, negative, or neutral.Advanced, "beyond polarity" mining classification searches, "angry”, “glad” and "sad”.Sentiment analysis is a natural language processing and information
i.e. Super Vector Machine (SVM) and Max margin Multi Label Classifier.

Research indicates that social media users may purposefully manage their online identity to “look better” than in real life [8], [10]. Other studies presents that there is a deficiency of awareness about managing online identity amongst students [11], and social media is regarded as their personal space by young people to hang out with peers outside the sight of parents and teachers [7]. Students’ online conversations reveal different experiences that are not easily seen in formal classroom settings, thus are usually not documented in educational literature. The abundance of social media data not only provides opportunities but also represents technological difficulties for analyzing large-scale informal textual data. The following section reviews popular methods used for analyzing Twitter data.

Jindal Rajni and Dutta Borah Malaya implemented a prediction analysis method that can help to improve the education quality in higher education for ensuring organization success at all level. They used the C5.0, C4.5-A2, C4.5-A1 algorithms for prediction analysis, after that they compare their results. The result of C5.0 is best in performance. Then they applied NN (Neural Network) and CRT
Algorithms on same data set for prediction analysis. After that they compared the result of C5.0 with Neural Network and CRT algorithms result. This paper analyzes the accuracy of algorithm in two ways; firstly is by comparing the result of C5.0 with C4.5-A2, and C4.5-A1. After that the C5.0 algorithm is comes out to be best algorithm in accuracy. Then its result is compared with NN (Neural Network) and CRT.

Ilkyu Ha et al. [5] presented an approach to mine sentiment data from several kinds of unstructured social media text data by utilizing parallel HDFS to secure social multimedia for sentiment analysisby applying MapReduce functions. The result indicates that proposed approach is close to the manual processes.

Mane, Sunil B et al. [6] presented a method which gives a technique of sentiment analysis utilizing hadoop which processes large amount on hadoop cluster quicker in real time. Twitter, the major and biggest social media site receives tweets in millions every day. This enormousquantity of raw data then can be utilized for industrial or business purpose by organizing according to our requirement and processing.

Rosa M. Carro et al. [7]in this paper, explained that researcher implemented this method in Sent Buk which is a Facebook application. The results acquired based on this method indicate that sentiment analysis in Facebook is possible with high accuracy of (83.27%).

3. Problem Formulation

The problem of sentiment analysis for meaningful information has attracted the interest of the researchers all around the globe for a while now. This thesis proposes to analyze social media data and draw inferences about the data. The problem is to extract the large unfiltered data of twitter as it forms a platform for extensive live data. Mining such data is a challenge in itself as they are largely unstructured and needs to be pre-processed. The next problem is to mine and extract important features from the large amount of unique extracted features. This is carriedfor the reduction of the dimensionality of the problem. Then scores of each words needs to be found out and finally opinions mining algorithms needs to be designed and implemented for extracting meaningful information. The classifiers needs to be analyzed in terms of precision, accuracy and recall for its performance.

4. Proposed Methodology

1. Date Extraction:

The twitter API named as 'tweepy' has been used in this thesis for the extraction step. The major steps involved in development of the framework for live streaming of tweets begin with setting up an account on twitter.

- Set up an account on twitter
- Go to dev.twitter.com
- Create a new app and register for it
- Change access level to Read, write and access messages
- Generate security id and secret number
- Generate access token id and secret token number
- Save them to be utilized for streaming

OAuth handler is used for streaming the tweets. Filters are applied on it using the track filter. The tweets are filtered by two ways:

- Filter by content
- Filter by location

Because to the policies of twitter the filtering is not absolutely correct and there might be a similar tweet which doesn't lie in the filtered bandwidth.

The location is done using a 'location' filter available with tweepy. The location filter works on the basis of longitude and latitude of the place. A bounding box has to be formed where the location filter works. Any tweets sent from that bounding box is streamed.Proposed system will consists of various modules. Each module incorporates diverse techniques to execute its definite tasks. When a specific module concludes its task, its result cum output will become input for the next module. Finally the collectivedetermination of each and every module will be displayed.

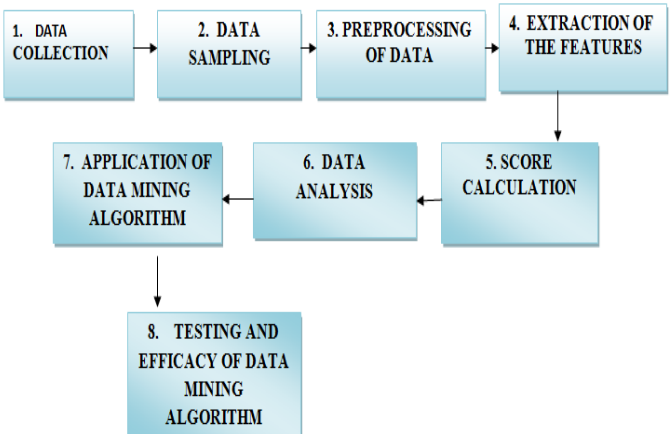
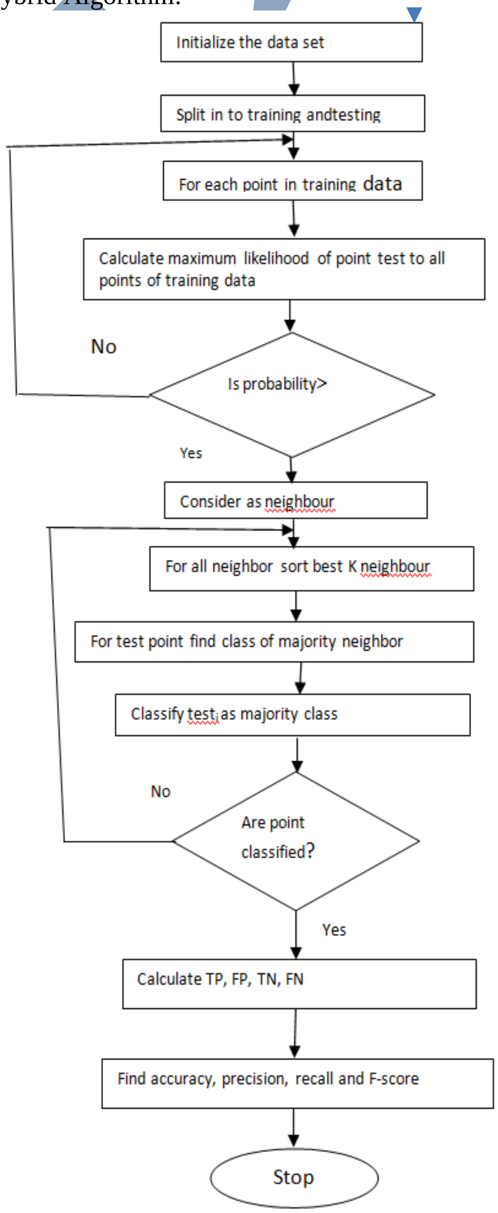

1. DATA COLLECTION
2. DATA SAMPLING
3. PREPROCESSING OF DATA
4. EXTRACTION OF THE FEATURES
5. SCORE CALCULATION
6. DATA ANALYSIS
7. APPLICATION OF DATA MINING ALGORITHM
8. TESTING AND EFFICACY OF DATA MINING ALGORITHM

Fig. 1 Block Diagram of Proposed System

The steps of implementation can be listed as:

1) The data regarding student problem is collected from social media site (Twitter).
2) Then sampling of the collected data is done through training and testing of data.
3) The preprocessing of text data is done i.e. streaming, tokenization, special character removal, stop work removal is done on text data.
4) In step 4 Extraction of features is done after preprocessing of the data.
5) In this step conversion of text data to numerical data is done using TFIDF () approach.
6) In this step data analysis is done.
7) Application of Data Mining algorithm is done.
8) In this step testing and efficacy of algorithms is done with previous used algorithm.

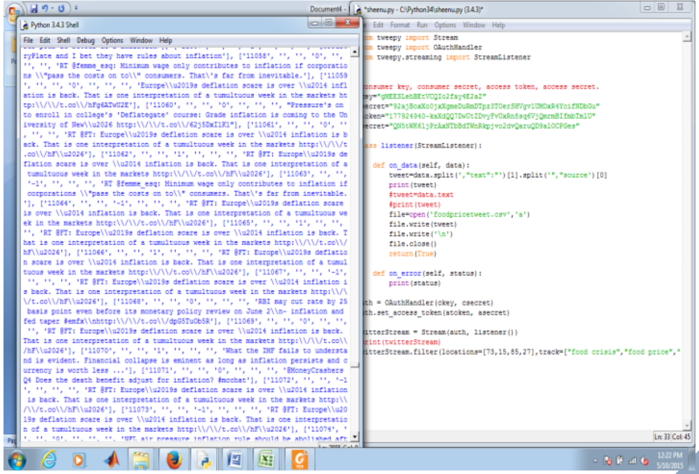
Hybrid Algorithm:


Initialize the data set
Split in to training andtesting
For each point in training data
Calculate maximum likelihood of point test to all points of training data
No
Is probability>
Yes
Consider as neighbour
For all neighbor sort best K neighbour
For test point find class of majority neighbor
Classify test as majority class
No
Are point classified?
Yes
Calculate TP, FP, TN, FN
Find accuracy, precision, recall and F-score
Stop

Flow Chart of Hybrid Algorithm

5. Results

These tweets are saved in a database and sentiment values are assigned to them based on manual interpretation. The sentiments are assigned as follows: '1' for positive sentiment, '2' for negative sentiment, '0 ' for neutral sentiment



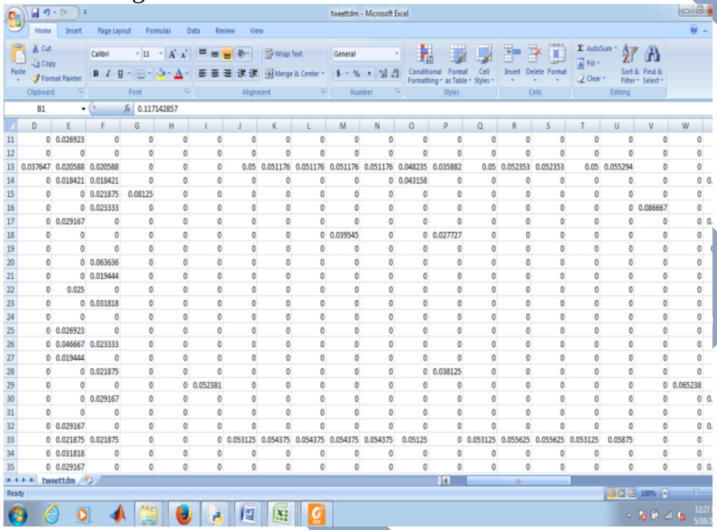
```
ryPlate and I bet they have rules about inflation'], ['11058', '', '', '0', '',
'', '', 'RT @femme_esq: Minimum wage only contributes to inflation if corporatio
ns \\"pass the costs on to\\" consumers. That\'s far from inevitable.'], ['11059
', '', '', '0', '', '', '', 'Europe\\u2019s deflation scare is over \\u2014 infl
ation is back. That is one interpretation of a tumultuous week in the markets ht
tp:\\/\\/t.co\\/hFg6ATwU2E'], ['11060', '', '', '0', '', '', '', "Pressure's on
to enroll in college's 'Deflategate' course: Grade inflation is coming to the Un
iversity of New\\u2026 http:\\/\\/t.co\\/62j5DxI1K1"], ['11061', '', '', '0', ''
, '', '', 'RT @FT: Europe\\u2019s deflation scare is over \\u2014 inflation is b
ack. That is one interpretation of a tumultuous week in the markets http:\\/\\/t
.co\\/hF\\u2026'], ['11062', '', '', '1', '', '', '', 'RT @FT: Europe\\u2019s de
flation scare is over \\u2014 inflation is back. That is one interpretation of a
 tumultuous week in the markets http:\\/\\/t.co\\/hF\\u2026'], ['11063', '', '',
 '-1', '', '', '', 'RT @femme_esq: Minimum wage only contributes to inflation if
 corporations \\"pass the costs on to\\" consumers. That\'s far from inevitable.
'], ['11064', '', '', '-1', '', '', '', 'RT @FT: Europe\\u2019s deflation scare
is over \\u2014 inflation is back. That is one interpretation of a tumultuous we
ek in the markets http:\\/\\/t.co\\/hF\\u2026'], ['11065', '', '', '1', '', '',
'', 'RT @FT: Europe\\u2019s deflation scare is over \\u2014 inflation is back. T
hat is one interpretation of a tumultuous week in the markets http:\\/\\/t.co\\/
hF\\u2026'], ['11066', '', '', '1', '', '', '', 'RT @FT: Europe\\u2019s deflatio
n scare is over \\u2014 inflation is back. That is one interpretation of a tumul
tuous week in the markets http:\\/\\/t.co\\/hF\\u2026'], ['11067', '', '', '-1',
 '', '', '', 'RT @FT: Europe\\u2019s deflation scare is over \\u2014 inflation i
s back. That is one interpretation of a tumultuous week in the markets http:\\/\
\/t.co\\/hF\\u2026'], ['11068', '', '', '0', '', '', '', 'RBI may cut rate by 25
 basis point even before its monetary policy review on June 2\\n- inflation and
fed taper #emfx\\nhttp:\\/\\/t.co\\/dpG5Tu0b5R'], ['11069', '', '', '0', '', '',
 '', 'RT @FT: Europe\\u2019s deflation scare is over \\u2014 inflation is back.
That is one interpretation of a tumultuous week in the markets http:\\/\\/t.co\\
/hF\\u2026'], ['11070', '', '', '1', '', '', '', 'What the IMF fails to understa
nd is evident. Financial collapse is eminent as long as inflation persists and c
urrency is worth less ...'], ['11071', '', '', '0', '', '', '', '@MoneyCrashers
Q4 Does the death benefit adjust for inflation? #mcchat'], ['11072', '', '', '-1
', '', '', '', 'RT @FT: Europe\\u2019s deflation scare is over \\u2014 inflation
 is back. That is one interpretation of a tumultuous week in the markets http:\\
/\\/t.co\\/hF\\u2026'], ['11073', '', '', '-1', '', '', '', 'RT @FT: Europe\\u20
19s deflation scare is over \\u2014 inflation is back. That is one interpretatio
n of a tumultuous week in the markets http:\\/\\/t.co\\/hF\\u2026'], ['11074', '
', '', '0', '', '', '', 'NFL air pressure inflation rule should be abolished aft
```

```
om tweepy import Stream
om tweepy import OAuthHandler
om tweepy.streaming import StreamListener


onsumer key, consumer secret, access token, access secret.
ey="gMEE5LehBEtVCQIo2fay4E2aZ"
ecret="92ajBoxXo0jxXgmeDuRmDTpz3TOerSHVgv1UMOxR4YcifNDbGu"
oken="177924940-kaXdQQ7DwUtZDvyFvUxRn8sq6VjQmrmBIfmbTmlU"
ecret="QN5tWK61jPrAxNTbBdTWnRkpjvo2dvQaruQD9alOCPGes"

ass listener(StreamListener):

    def on_data(self, data):
        tweet=data.split(',"text":"')[1].split('","source')[0]
        print(tweet)
        #tweet=data.text
        #print(tweet)
        file=open('foodpricetweet.csv','a')
        file.write(tweet)
        file.write('\n')
        file.close()
        return(True)

    def on_error(self, status):
        print(status)

th = OAuthHandler(ckey, csecret)
th.set_access_token(atoken, asecret)

itterStream = Stream(auth, listener())
rint(twitterStream)
itterStream.filter(locations=[73,15,85,27],track=["food crisis","food price","
```

Python 3.4.3 Shell
File Edit Shell Debug Options Window Help
sheenu.py - C:\Python34\sheenu.py (3.4.3)
Edit Format Run Options Window Help
Ln: 33 Col: 45
12:22 PM

Fig.3 Sentiments Assigned

An array of the tweets is created and term document matrix is created using TFIDF score as shown below.



B1 0.117142857

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
11	0	0.026923	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0.037647	0.020588	0.020588	0	0	0	0.05	0.051176	0.051176	0.051176	0.051176	0.048235	0.035882	0.05	0.052353	0.052353	0.05	0.055294	0	0
14	0	0.018421	0.018421	0	0	0	0	0	0	0	0	0.043158	0	0	0	0	0	0	0	0
15	0	0	0.021875	0.08125	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0.023333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.086667	0
17	0	0.029167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0.039545	0	0	0.027727	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0.063636	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0.019444	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0.025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0.031818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0.026923	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0.046667	0.023333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0.019444	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0.021875	0	0	0	0	0	0	0	0	0	0.038125	0	0	0	0	0	0	0
29	0	0	0	0	0	0.052381	0	0	0	0	0	0	0	0	0	0	0	0	0	0.065238
30	0	0	0.029167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0.029167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0.021875	0.021875	0	0	0	0.053125	0.054375	0.054375	0.054375	0.054375	0.05125	0	0.053125	0.055625	0.055625	0.053125	0.05875	0	0
34	0	0.031818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0.029167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 4 Creation of Term Document Matrix using TFIDF

Hybrid algorithm and Naïve Bayes algorithm is applied on the data and the results are obtained.

Training to test ratio is kept as 3:1. A total tweets is finally selected after filtering and all and manual assignment of sentiments is done to be fed into the classifier. A hybrid Naïve Baye's –KNN is implemented and the results are compared as below.

5.1 Performance metrics

Performance of selected classifier for our work is compared from the given metric described below:**i. Precision:** Precision is calculated as percentage of examples expected as belong to class x that is actually correctly predicted. This is defined as:

$$Precision(x) = \left(\frac{number\ \ of\ correctly\ \ classified\ \ instances\ \ of\ class\ x}{number\ \ of\ instances\ \ classified\ \ as\ belonging\ \ to\ class\ x}\right) \times 100$$

ii.Accuracy: Accuracy is calculated as fraction of sum of correct classification to total number of classification. It is defined as:

$$Accuracy(x) = \left(\frac{sum\ of\ correct\ \ classification}{total\ number\ \ of\ classification}\right) \times 100$$

The Confusion matrix are as follows:

TP=30

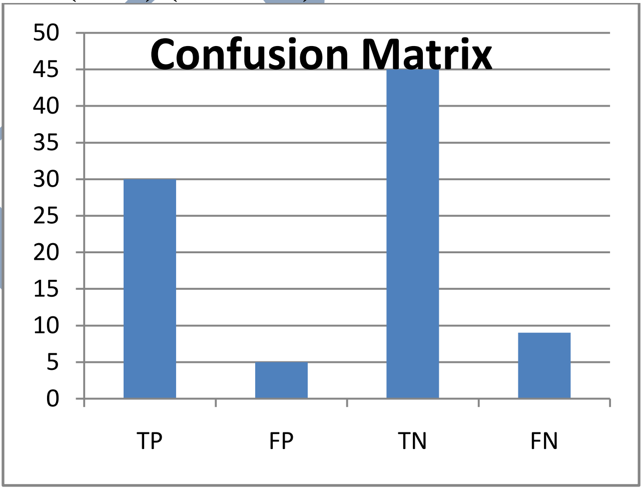
FP=5

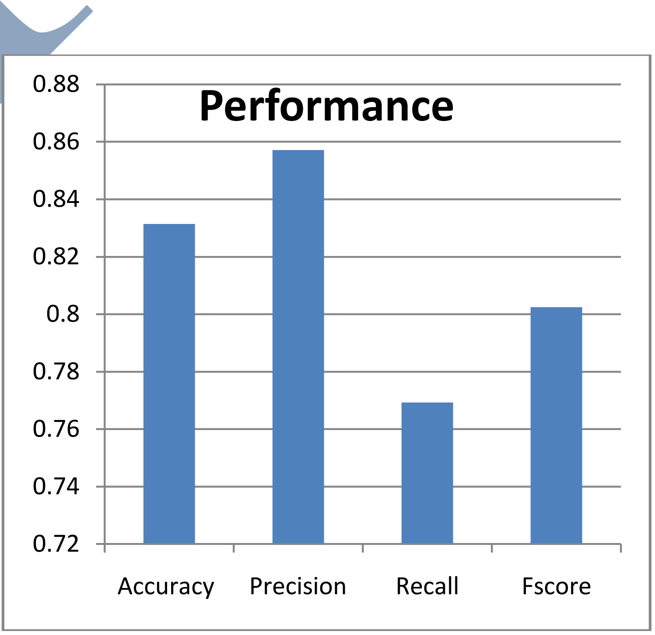
TN=45

FN=9

The accuracy comes out to be

Acc= (30+45)/ (30+45+5+9)*100= 84%





6. Conclusion

A methodology for the classification of sentiments was developed in this thesis for educational data mining crisis in Indian market. Twitter API was used for streaming of tweets. Stemming was done to all words for extract the root words.TF-

IDF score based approach was utilized and the score was calculated for each tweets. Feature Selection was applied on it using Chi Square method and information gain. The extracted features form a term document matrix which is utilized in the classification algorithm. Two classification algorithms are compared as presented in previous chapter.The results are found to be satisfactory and when comparative analysis is done between them it is found that hybrid performs better.

References

[1]. V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan,“Towards building large-scale distributed systems for Twittersentiment analysis,” in Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC ’12), pp. 459–464, March 2012.

[2]. Horakova, Marketa. "Sentiment Analysis Tool using Machine Learning." *Global Journal on Technology* (2015).

[3]. Bames New Approach to the Design of Decision Support System to Improve E-Leaming Environments, 26–29,The 4th International Conference on E-learning and E-teaching (ICELET) 2013.

[4]. Jebaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." *International Journal of Computer Applications* 47.11 (2012).

[5]. Ilkyu Ha, Bonghyun Back, and ByoungchulAhn,"MapReduce functions to analyze sentiment information from social big data" International Journal of Distributed Sensor Networks.Id417502,2014

[6]. Mane, Sunil B., YashwantSawant, SaifKazi, and VaibhavShinde. "Real Time Sentiment Analysis of Twitter Data Using Hadoop." *International Journal of Computer Science and Information Technologies,(3098-3100)* 5, no. 3 (2014).

[7]. Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." *Computers in Human Behavior* 31 (2014): 527-541.

[8]. EfthymiosKouloumpis , Theresa Wilson, and Johanna Moore, “Twitter sentiment analysis: the god the bad and the OMG!,” in Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 538–541, 2011.

[9]. Flume, http://flume.apache.org/ [10]. G.Vinodhini, RM.Chandrasekaran. "Sentiment analysis and opinion Mining: A Survey " , Volume 2, Issue 6,International Journal of Advanced Research in Computer Science and Software Engineering.2012.