iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Mining Twitter Data of Rural Area Engineering Colleges for Understanding Issues and Problems in Their Educational Experiences

S. D. Rane[1], U. A. Nuli[2], N. S. Mahajan[3]
[1,2,3] *DKTE's Textile and Engineering Institute, Shivaji University, Kolhapur,India*

Abstract*: Students' informal conversations on social media such as Twitter are useful for understand their learning experiences, and feelings. Data from such social media environments can provide valuable information about students learning system. Collecting and analyzing data from such media can be difficult task. However, the large scale of data required automatic data analysis techniques for classify twitter data . We developed new system to combination of qualitative analysis and large-scale data mining techniques. This system focuses on engineering students' Twitter posts which are collected from rural area engineering colleges to understand issues and problems in their learning. First we conduct a qualitative analysis on tweets collected from engineering colleges using term #DStudentsproblems. Collected tweets are related to engineering students' college life. In proposed system we used a multi-label classification algorithm to classify tweets reflecting students' problems such as soft skill issues, heavy study load, lack of social engagement, and sleep problems.
Keywords*: Data mining, social networking, tweet analysis, classification.

I. INTRODUCTION

Twitter have become important venues for the engineering student's to communicate and exchange information about their daily learning process. Students also discuss and share their everyday problems in an informal manner on twitter and Whatapps. Students' written information on twitter provide implicit knowledge and a whole new perspective for educational researchers to understand students' learning experiences outside the controlled classroom environment. This understanding can be useful for improvement of education quality in college and enhance student recruitment and retention ratio.

Traditionally, educational researchers uses techniques such as surveys, student's interviews, classroom activities such as take student's feedback related to students' learning experiences. These methods are usually required more time as compare with automate large scale data mining technique. Students are given the feedbacks under the pressures of faculties. The scales of such studies are usually limited to particular college or class.

Proposed system give more focuses on rural area engineering college students' posts on Twitter about problems in their educational experiences because Engineering colleges and departments have been struggling with student recruitment and retention issues, Based on understanding of issues and problems in students' life, policymakers, management can make decisions on services that can help students overcome such problems and issues. Propose system focused on a workflow for a qualitative research methodology and large-scale data mining techniques . We use qualitative data from human interpretation for data mining algorithm, so that we can gain deeper understanding of twitter data and get quality result based on category defined during content analysis process. The objectives of proposed systems are 1)Developed module for extract twitter data using twitter API 2) To developed tweets cleaning module for remove noise from tweets and tweets classification module for tweets classification using predefined categories using content analysis.

II. PROPOSED SYSTEM

In the proposed system we focused on rural area engineering colleges students' Twitter posts to understand issues and problems in their educational experiences. First we Extract quality twitter data using a input terms #DStudentsproblems, #AluminiSuggestions and #engineeringProblemst, then later perform content analysis on collect data for category formation. Use a multi-label classification algorithm to classify tweets reflecting students' problems such as soft skill issues, heavy study load, lack of social engagement, and sleep problems. The proposed scheme is made up of twitter data extraction, tweets data cleaning. Classification of tweet data and web module .The The proposed scheme performs various operations on tweets as shown in figure1.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

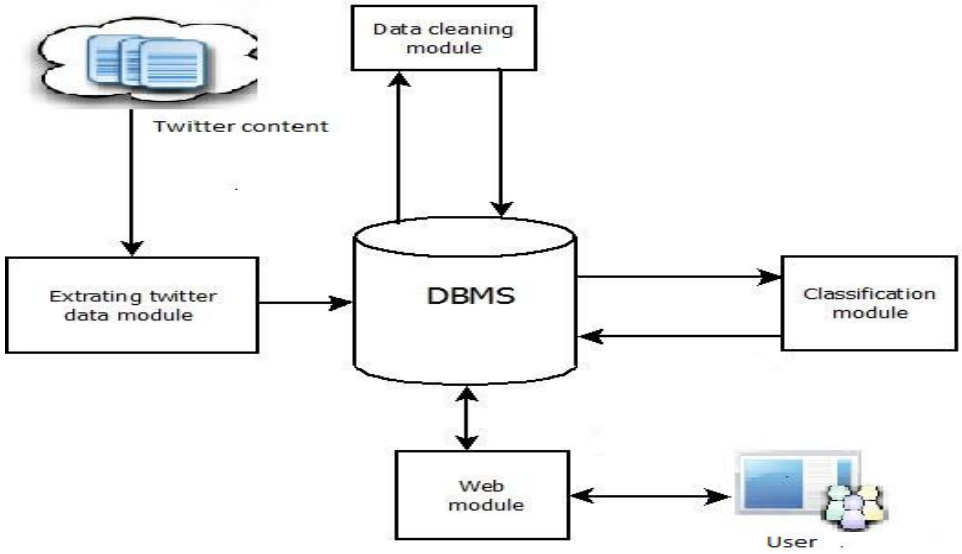


Fig.1 Architecture of Proposed System

In the first phase user extract tweets from twitter using twitter standard API[10]. Tweet processing operation performed in second phase. Then, tweet classification is perform using Naïve Bayes algorithm, tweets are classified into heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other. In data cleaning phase perform various operation on tweet to remove noise from it as shown in Fig 2.

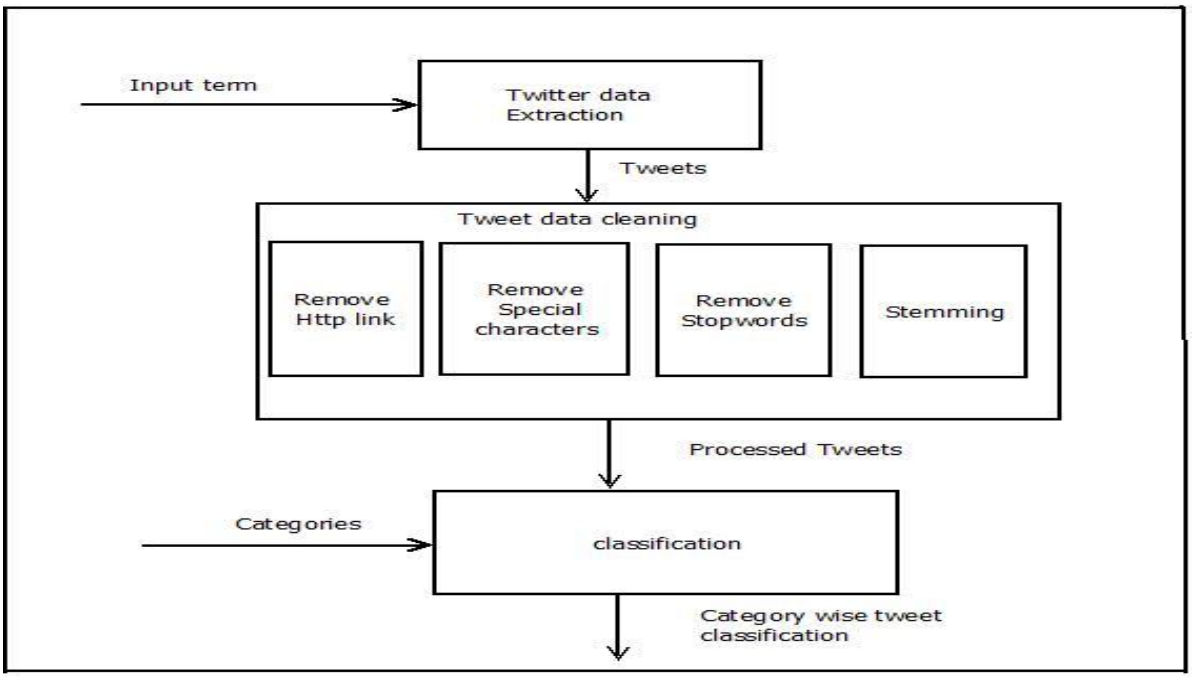


Fig. 2 Block Diagram of Proposed System

A. Extracting Twitter Data

To extract twitter data first go to official Twitter Developer Registration at https://dev.twitter.com/apps/new and register application detail there like application name, Description. We have created new application "Mining Social media data". Twitter account is required for application registration. After registration completion we visit the settings page and adjust the setting for application access "Read, Write and Access direct messages" to get our application the full functionality. Need to bind Twitter account with the application we registered. Once we finish the binding process, we get the keys and tokens (i.e., a pair of consumer key and

consumer secret and a pair of access token and access token secret) for our application. We have taken query term "engineeringProblems",AluminiSuggestions and DStudentsproblems as input to collect tweets data, In result we get tweets related to query term.

B. Tweet Data Cleaning and Text Pre-processing

In this module we pre-processed the texts and find useful text before training the classifier. We take input as tweets which are collect using 1st module. Perform cleaning operation on tweets because there is noise present in to collected tweets so there to Pre-processing the tweets before training the classifier.

Use MySQL5.5 database for store tweets. First create database "social_data_mine". Using create database social_data_mine command in sql. Under the "social_data_mine" database create various tables, which are required for store data. Tables are tweet, search_topic and processing_tweet .Tweet table is used to store the collected tweets. Un-processed tweets are store in to tweet table. search_topic table is used to store search topic name(query term) and topic id and processing_tweet table is used to store processed tweet .

1) Remove Http Link **:** There is no use of http link for tweet classification, so remove the http link from the collected tweets.

2) Remove Special Characters: Removed all the #DStudentsproblems hashtags. For other co-occurring hashtags, Only removed the # sign, and kept the hashtag texts, Removed all words that contain non-letter[0-9] symbols and punctuation. such as # ,:$,% ,?,/,>,=,!,|,(etc Scan the processed tweet message, if the any special character found in the tweet message then it replace with the blank and return the tweet without special character.

3) Stemming : Stemming means reducing a word to its base (or stem).Stemming is useful when doing any kind of text analysis concerned about the content of a the different times of verbs and the different ending for singular and plural , make it difficult to discern the importance of specific words with in text when treat each word as it is .Use a dictionary that lists all words together with their stems. wordnet dictionary is large lexical database of English Nouns, Verb, Adjectives and Adverbs..

Scan the input processed tweet message and make the token from the tweet message and check a baseform of token. Find the word match do with help of LookupBaseForm().Get the wordstem form index word, index word used for organize the word in wordnet dictionary. After stemming get processed tweets.

4) Remove Stopwords : Remove the words that are very commonly used in a given tweets, because to focus on the importance words. Tweet text contain stop words such as hi, etc, be, as and many more words. After processing all tweets, it store into processing_tweet table.

C. Tweets Classification

We used multi label Naïve bayes classifier to classified tweets based on categories. Take input as processed tweets. Apply classification algorithm on processed tweets for categories wise classification of tweets. Tweets are classifieds into prominent categories such as heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other. Algorithm result is store into Navebayes table.

1) Categories Development: There were no pre-defined categories of the data so need to explore what students were saying in the tweets. Perform inductive content analysis for categories development. Inductive content analysis is to identify what are the major worries concerns and issues that engineering student encounter in their study and life. Use inductive content analysis[3] for categories development. Prominent categories are heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other.

2) Naïve Bayes multi-label classifier: Basic procedure for multi-label classifier. Each tweet is considered as document, there are a total number of N words in the learning dataset tweets collection $W=\{w_1, w_2,\ldots,w_N\}$ and total number of L categories $C=\{c_1,c_2,\ldots\ldots c_L\}$.Learning dataset contain most probable words for each category. Learning dataset is key words based dataset.

Suppose there total number of M words in the training set and A of them are in category c. Then the prior probability of category c is

$$P(c) = \frac{A}{M}$$

And prior probability of other category is

$$p(c') = \frac{M-A}{M}$$

Suppose total number of Wn words are appears in category c from selected tweet and category have Cn words , based on likelihood estimation, the likelihood probability for category c is

$$\text{Likelihood probability for category c} = \frac{Wn}{Cn}$$

For a tweet Ti in the dataset set, there are k words $W_{di} = \{w_{i1}, w_{i2}, \ldots, w_{ik}\}$ and W_{di} is subset of W. The purpose is to classify this document into c or not c. We assume independence among each word in this documents, any word w_{ik} condition on c or c' follows multinomial distribution .Therefore the posterior probability of d_i belongs to category c is

$$\text{Posterior probability } P(c|T_i) = \frac{\text{prior probability for category c}}{\text{likelihood probability for category c}}$$

If p(c|Ti) is larger than the probability threshold t, then di fit into category c, otherwise, di does fit into category c'. Other is only category if posterior probability of all categories are less than threshold value. Result of Navie bayes classifier is stored in to table.

3) Navie Bayes Clasifier Algorithm : This algorithm considers each sub words in the review and accordingly classifies the reviews in different categories

Let T is the Tweet

Step 1: Define categories $C = \{c_1, c_2, \ldots\ldots c_L\}$.
Define categories such as heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other

Step 2: Find prior probability for each category

Step 3: Read tweets from a database.

Step 4: Divide T into sub words $\{w_1, w_2, \ldots, w_N\}$

Step 5: Check sub words $\{w_1, w_2, \ldots, w_N\}$ for every categories

Step 5: If words match with categories $\{c_1, c_2, \ldots\ldots c_L\}$. Increment the counter for those categories

Step 6: Find the likelihood probability for each category

Step 7: Find the posterior probability P(c|T) for each category
If P(c|T) is larger than the probability threshold t, then T fit into category c, otherwise, T does fit into category c'.

D. Web module

This module shows the classification result to users or policy makers. We created form for registration, login, tweets result, and category wise result. User gives the input as user name and passwords for the login to this module, after login user select tweet result dataset for category wise result. Category wise result form shows the result according selected category. Based on result, the Policy makers can take decision for improve education quality in college and enhance student recruitment and retention ratio.

III. RESULTS AND PERFORMANCE MEASURES

A. Experimental Dataset

It is challenging task to collect students learning experience data from twitter because irregularity of students post data on twitter. We Extract data using input terms DStudentsproblems, AluminiSuggestions, and engineeringProblems . This are the most popular hashtag specific to collect rural area engineering colleges students twitter post. In total, we collect 20000 tweets using twitter API., input term hashtag DStudentsproblems, AluminiSuggestions, and engineeringProblems.

B. Performance Measures

A commonly used measure to evaluate the performance of classification algorithms includes accuracy, precision, recall. We used label base measure for performance evalution. Label based measures are calculated based on predefined each category and then average over all category. Create matrix for corresponding category c

TABLE I : Matrix for Corresponding Category c

	True select tweets for c	True not select for c
Expected Tweet	True Positive(TRP)	False Negative(FAN)
Not Expected Tweet	False Positive(FAP)	True Negative(TRN)

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

The sum of TRP,FAN,FAP and TRN equal to total number of tweet. Then for this one category

Accuracy a=$\frac{TRP+TRN}{TRP+TRN+FAP+FAN}$,

Precision p=$\frac{TRP}{TRP+FAP}$,

Recall r=$\frac{TRP}{TRP+FAN}$.

TABLE II: Label Base Accuracy and Precision

Category	Label a.	Label p.
Heavy study load	0.881	0.883
Soft skill issues	0.971	0.976
Negative emotion	0.914	0.915
Sleep problems	0.897	0.898
Lack of social engagement	0.933	0.935
Other	0.779	0.780

Result for label base accuracy and precision for Naive Bayes classification algorithm for various category has shown in Table

C. Classification Results

In category development stage, we had total of 1000 #DStudentsproblems ,#AluminiSuggestion and #engineeringProblems tweets annotated with 6 categories. We used 75% of the 1000 tweets for training (750) and 25% for testing (250).We used learning words dataset .In this dataset contain 30 probable word for each category .We apply training model on rest of 20000 tweets and found total 1100 tweets reflecting the five categories of tweets classification. One tweet may be fall several different categories, so sum of tweets for all categories more than 1100 as shown in fig.3.

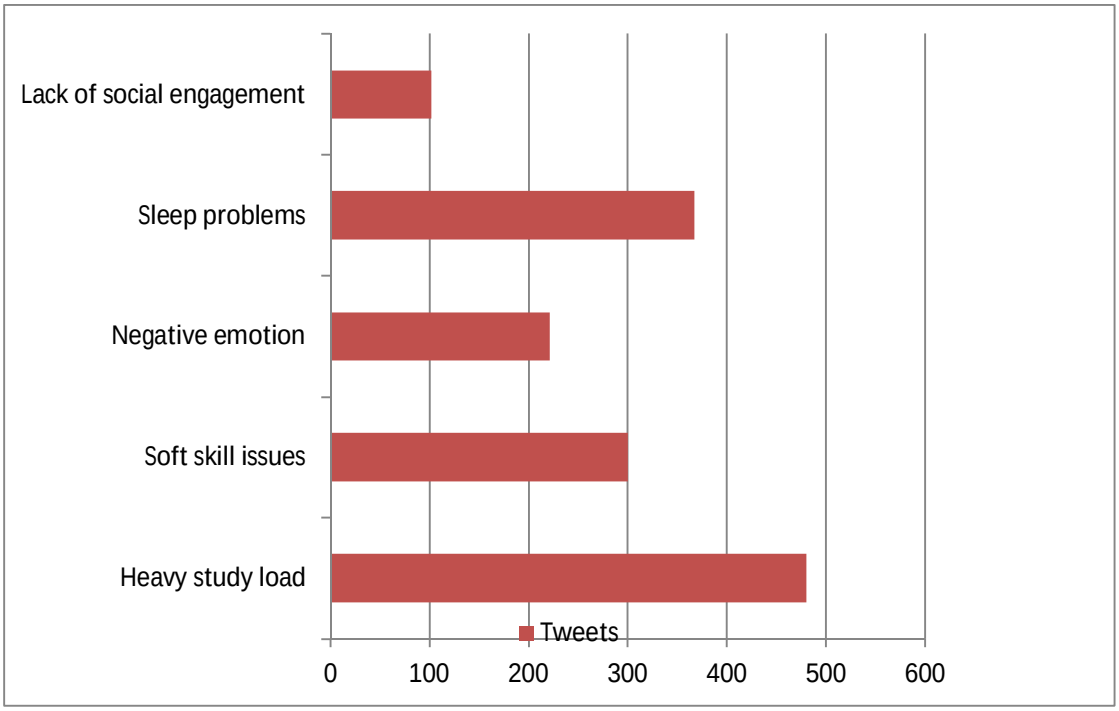


Fig.3 No of tweets for each category determine from dataset

Fig.4 shows category wise result for Nnaive Bayes classifiers, result contain sr_no, tweet text and categories of tweets. One tweet may fall into multiple category.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

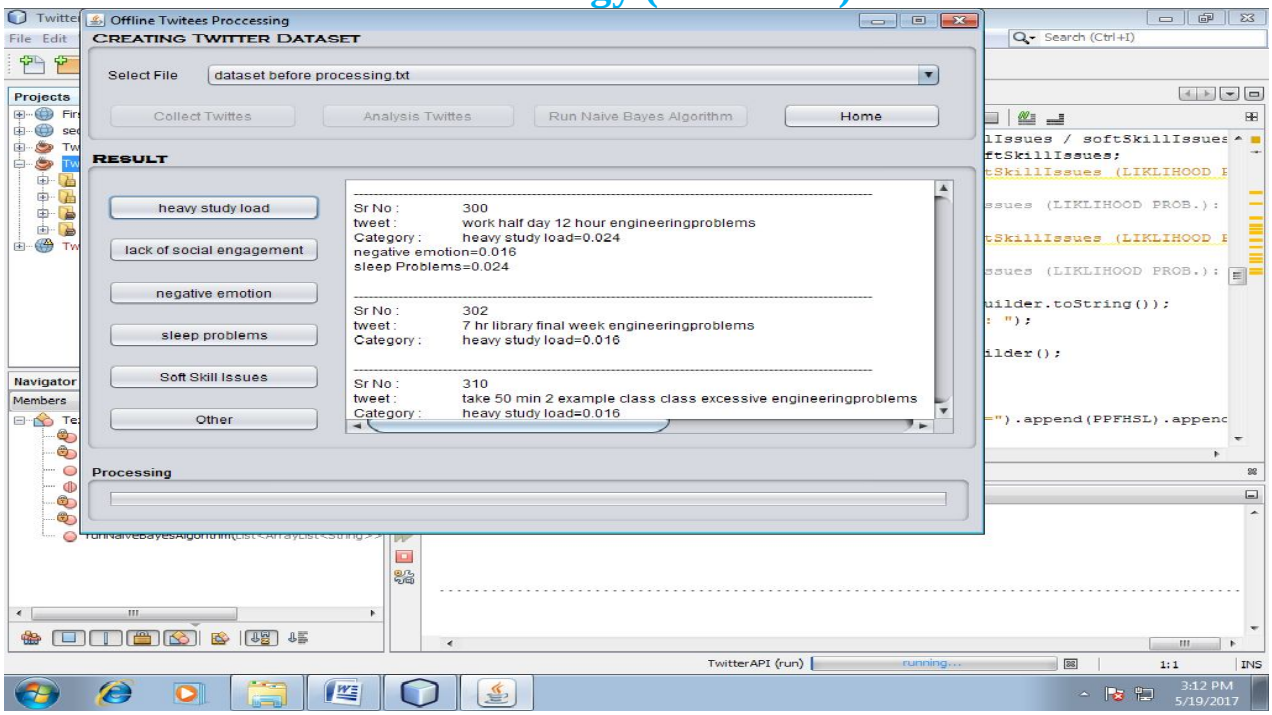


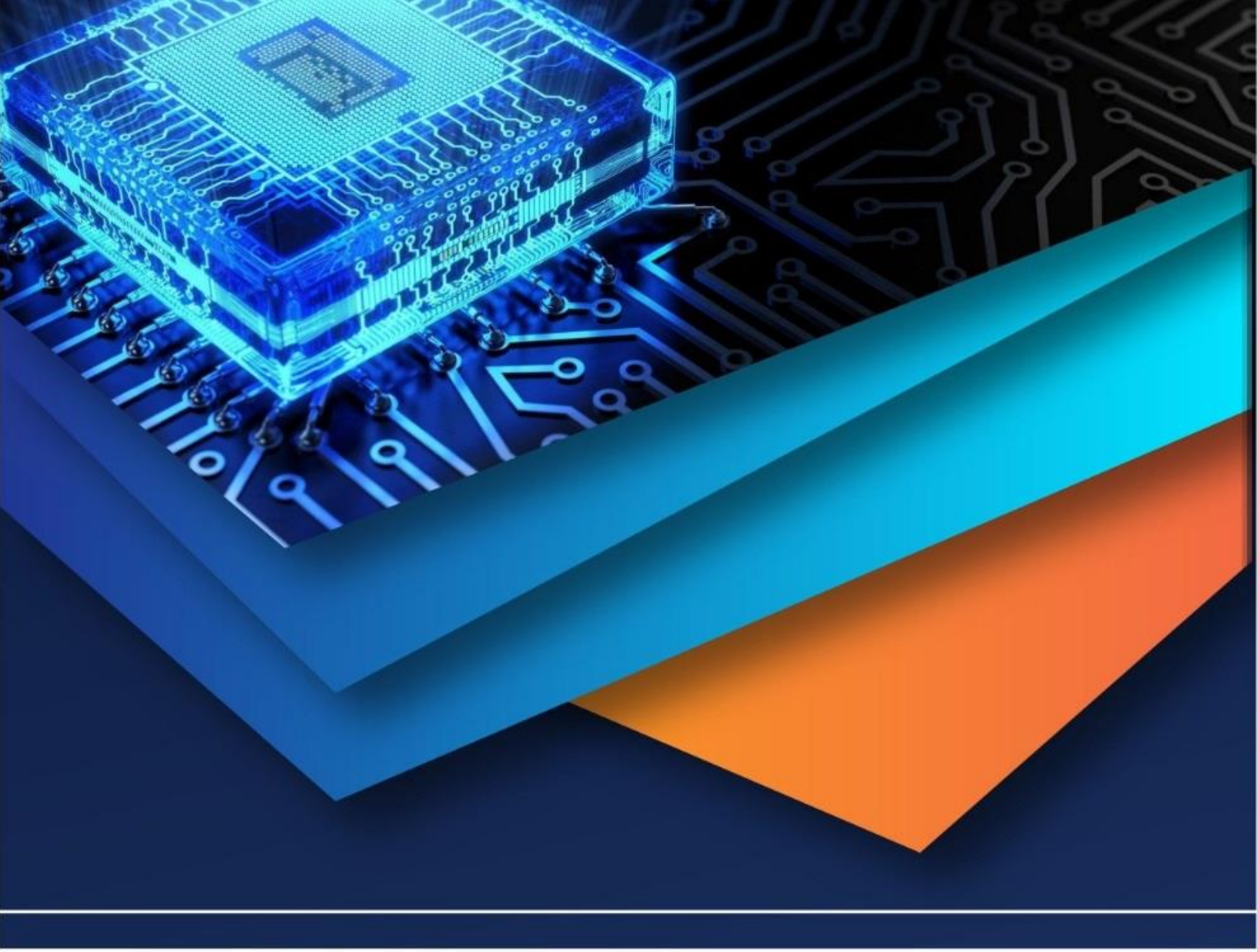
Fig.4 Naïve Bayes classification algorithm Result for selected dataset

IV. CONCLUSION

This research presents data mining techniques and analysis about rural area engineering colleges' students learning experience using twitter tweets data. Students' tweets are follows in various categories. In this study addressed main problem of rural area engineering students such as soft skill issues. Our study inform educational policy maker to gain understanding of engineering students colleges problems. For future work could analyse problems of engineering faculty related to college using social media such as twitter and Whatsapps.

REFERENCES

[1] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representation and communication: challenges in interpreting large social media datasets," in Proceedings of the 2013 conference on Computer supported cooperative work, 2013.
[2] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 241–249.
[3] Xin Chen, Mihaela Vorvoeanu, and Krishna Madhavan, "Mining Social Media Data for Understanding students learning Experience, IEEE Transactions 2014.
[4] A.Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, pp. 1–12, 2009.
[5] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 23–32.
[6] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in Proceedings of the 20th international conference companion on World wide web, 2011, pp. 57–58.
[7] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in Proceedings of the 33rd nternational ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2010, pp. 841–842.
[8] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, pp. 601–618, 2010.
[9] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," Data mining and knowledge discovery handbook, pp. 667– 685, 2010.
[10] "Using the Twitter Search API Twitter Developer" [Online]. Available:https://dev.twitter.com/docs/using-search.[Accessed:11-May-2013].



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY