Understanding Students' Learning Experiences By Mining Social Media Data

Garima Chanana, V.Vijaya Kumar, M.Geetha

Abstract— Conversations of students on social networking sites like twitter, facebook throw light on education experiences like emotions, concerns. Twitter is a micro-blog with each tweet within 100-150 words so we can understand emotions of candidates. Most tweets are related to emotions, which tweets fall under which emotion. In this paper we are focusing to develop a model which predicts student's emotion and understand their feelings, opinions related to their educational experiences. Few labels which we have used for fetching the tweets related to students are exams, results, engineering. Main phases in this application are text cleaning, processing, validation and prediction. In pre-processing /cleaning phase stop-words removal, stripping white-space, removing punctuation. In processing phase, document term matrix, creating corpus and applying supervised learning paradigms on training data. We validated the accuracy of model using 5-fold cross validation in validation phase. On the basis of training data, predicted the label of tweets in test data.

Keywords—social media, college, twitter, tweets , text mining ,supervised learning , machine learning , visualization ,preprocessing , classification , SVM , sentiment analysis.

I. INTRODUCTION

Social Networking sites like facebook, twitter are used for users to communicate with each other without caring for the ethical values. They help in enriching our knowledge and sharing it with no regards to distance, language and time gap.

Thus provides a platform for people from all over the globe to interact and engage themselves in groups for discussions that suite their requirements.

Understudies' computerized impressions give tremendous sum of verifiable learning and a radical new point of view for instructive specialists and experts to get it understudies' encounters outside the controlled classroom condition. This comprehension can educate institutional basic leadership on intercessions for in danger understudies, change of training quality, and subsequently upgrade understudy enlistment, maintenance, and achievement. The wealth of web-based social networking information gives chances to comprehend understudies' encounters, yet in addition raises methodological troubles in comprehending online networking information for instructive purposes. Simply envision the sheer information volumes, the decent variety of Internet slang, the capriciousness of area furthermore, timing of understudies posting on the web, and additionally the many-sided quality of understudies' encounters.

Social media site provides a great venue for students to share their emotions and different experiences. It lets students discuss their day to day experiences in casual way. This can help educationists to understand the student's behavior outside the class environment. This can help us in understanding the student's experiences and help institutions in decision making for at-risk students, improving quality of education, enhancing retention of students and success.

Twitter being one of the most popular social networking site which has shown a rapid growth in few years since its creation. There are more than 200million users who create more than 300million tweets per day. Most of the tweets are related to emotions that what tweets fall under which kind of emotion. Data on twitter is completely unstructured as people don't care about the grammatical structure and format of sentences while posting their tweets. With this large amount of data on twitter, extracting logical patterns from the data with correct information from this unstructured data is a tedious task to be performed.

Text mining can help to overcome this problem as it provides us with intelligence in computation in multiple disciplines like information retrieval, Statistics, Artificial intelligence and other fields.

Text mining is used for processing of unstructured(textual) data, extracting meaningful information from the text and providing the mined text available to various data mining algorithms. This information will help us to derive the summary of the tweets based on the words contained in them.

This will help us to analyze the words, cluster of words and determine the relationship between them and how they are related to other variables of interest.

Based on the problems and the issues mentioned above we are trying to extract the tweets from twitter and classify the tweets in test data based on different labels that we take in training data like engineering, exams, results, college, student etc. In Classification we classify text to a suitable category. Here there are four main phases to extract the data from the source and convert it into an understandable format for further use.

Revised Version Manuscript Received on March, 25, 2019.

Garima Chanana, VIT University Chennai, Tamilnadu, India.
Dr.V.VijayaKumar, Associate Dean, VIT University Chennai, Tamilnadu, India.
M.Geetha, Assistant Professor, VIT University, Chennai, Tamilnadu, India.(E-Mail: geetha.m@vit.ac.in)

II. LITERATURE SURVEY

A) Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse G¨oker, Yiannis Kompatsiaris, Alejandro Jaimes[4]

Online social networking sites and news websites provide updated and rich information about all kind of events that are happening on the real world at that moment. Detecting the trending topics is one of the basic requirement to analyze and condense the information generated from social media. Here they are comparing six topic detection techniques on few twitter datasets that are related to significant events , which differ in scale of timing and rate of attrition of the topic. We observed that how the volume of activities , procedures for sampling, data pre processing and the event's nature which we are considering, how these things greatly affect the topic detection ,moreover this thing also depends on the technique or method that is being used for the detection of topic.

B) Jie Yin, CSIRO ICT Centre Andrew Lampert, Palantir Technologies[7]

Situation awareness is "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future."

The system described in this paper uses NLP(natural language processing) and the techniques for data miningso that they can extract the information of the situation awareness from the twitter messages that are produced during various crisis and natural calamities.

C) Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu[10]

This paper is focusing on the problems in mining the different social emotions from the textual conversations with the rapid growth of web 2.0, plenty of documents are held responsible by social users with emotional labels or tags like happiness, sadness, surprise, shocked. Such kind of emotions can provide an new way for categorizing of the documents and therefore help the networked users to select the particular documents on the basis of their emotional priorities. Their motive is to get identify the connections between different things like emotions and effective terms and on the basis of this they can predict the social emotion from the textual data accordingly.

D) Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, Member, IEEE, and Francesco Marcelloni, Member, IEEE[9]

In this paper, they have represented a real-time monitoring system for event detection of traffic from stream analysis of twitter. In this, it fetches the tweets from Twitter according to several filters; then processing of tweets,applying techniques to mine the text; and then at last performing classification of the tweets. The motive is to allocate an appropriate class label to each tweet whether that tweets is related to traffic or not. Here in this paper we are monitoring the different areas of Italian Road, and permitting the detection of traffic in the real time. They have used SVM(support Vector Machines) model for the classification the tweets and they have achieved a high level of accuracy by solving the problem of classification (tweets related to traffic or tweets not related to traffic).

E) Donald E. Brown, Fellow, IEEE[8]

In this paper they have described the use of text mining with the union of techniques to detect the characteristics of accident

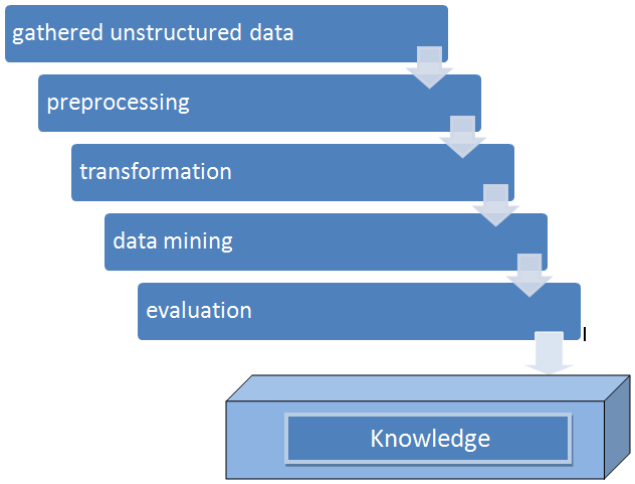

Fig.1. Block Diagram to explain the steps in text mining.

that can give us a clear picture of the things which have contributed for the accidents. In this they have evaluated the efficiency of text mining of the narrators of the accidents by determining the predictive performance of the extreme accidents value. The results however show that the accidents costs determined by the predictive accuracy has improved a lot with the use of the features that are there in text mining and this method of accuracy also improves the use of the latest ensemble methods.

F) Yuya Shibua ,IEEE 2017[10]

Online networking information, from Twitter and Facebook, for instance, can be viewed as basic data sources amid calamities through their utilization in identifying and surveying fiasco circumstances. This investigation outlines pertinent writing from the point of view of online networking for fiasco administration. The discoveries of this investigation demonstrate that while numerous past examinations have concentrated on the best way to use online networking information for moderating and reacting to calamities, few have concentrated via web-based networking media use for a catastrophe struck group's recuperation. This paper additionally contends that there is a need to contemplate the connections between's online networking information and the influenced individuals' recuperation exercises in reality. On account of this hole, the creator talks about one potential road for future work.

G) E Rejeesh , M Anupama , IEEE[11]

The focal point of this examination was to utilize web-based social networking and information mining

empowered pre-guiding session strategy to liven up the adequacy of separation training advising sessions. This activity is valuable for ideal use of advising sessions among

separate instruction students. 19 understudies and 2 advisors from four BCA groups of Indira Gandhi National Open University (IGNOU) were examined for the investigation. The understudy's and guides had been incorporated utilizing WhatsApp application. IGNOU gives directing sessions to elucidate questions of the students. They are utilizing two phase framework to assess the student. This exploration examine presented WhatsApp empowered pre-guiding session, target write question based test component and choice tree based examination before the advising sessions. Test answers gathered through WhatsApp are funneled to information mining process and attracted the choice tree to discover learning handicaps. In this examination we recognized innovation empowered pre-directing sessions enhanced the viability of guiding sessions.

III. EXISTING APPROACH

In the existing system the people are focusing on the issues which are prominent and ignoring the ones which are their in the long tail(do not have much importance). But these issues are also large in number because we need to analyze all the problems which students are suffering from so that we can find solution for the problems. They focused only on the problems which are being faced rather than both the positive and negative aspects. There is no use of NLP techniques for processing the data.

IV. PROPOSED WORK

Here we are getting the tweets related to educational data base on some keywords like engineering, classes, results, exams, college. Based on this we will perform sentiment analysis that which type of tweet fall under which count of emotion based on their counts.

After that creating the training data having more than 3000 tweets with the labels associated to them. Based on the training data we have tried to predict that which tweet in test data fall under which class label. After this validation comparison of the performance of different supervised learning paradigms.

V. TEXT MINING

In this section we are going to explain about the high level implementation architecture of text mining.

Figure 1 shows the various phases of the text mining process.

A. Definition of Text Mining

Text mining, additionally alluded to as text data mining, generally equal to content examination, is the way toward getting superb data from content. Top notch data is normally determined through the formulating of examples and patterns through means, for example, factual example learning. Content mining more often than not includes the way toward organizing the information content (normally parsing, alongside the expansion of some inferred semantic highlights and the evacuation of others, and resulting addition into a database), determining designs inside the organized information, lastly assessment and understanding of the yield. 'High caliber' in content mining more often than not alludes to some mix of significance, curiosity, and intriguing quality.Its main focus is on text categorization , grouping of the texts as well as for extracting and reducing the texts.

B. Phases in Text Mining:

- Data selection: The data needs to be selected from the collection of huge amount of data before extracting the information in the knowledge discovery in the databases. Data obtained is stored separately from the operational DBs.
- Pre-Processing/cleaning: In this phase we need to remove the inconsistent data, duplicate words, and correcting errors if any.
- Transformation: It involves the conversion of data from one format to another i.e. from the format of source file system to the required destination file .
- Data Mining: It involves finding useful patterns and information from the data using varied techniques or methods.
- Evaluation/Interpretation: Patterns generated from the data mining process needs to be represented in the form that is comprehensive so that it can be understood by everyone. This phase is used to check whether the pattern generated from this doesn't collide with the previously known facts.

C. Classification:

There are various algorithms under data mining like supervised, unsupervised and semi-supervised learning. Classification comes under the category of supervised paradigm where the dataset is split into two parts i.e. the training and test parts which has known class labels. Classification provides the records in which we know output of interest and then algorithm ahs to learn and predict the value for the new records where we don't have the output.

D. Knowledge:

At the end we get the predicted labels for the test data. Now we have the knowledge about the records and the labels associated with them. From this we can interpret the data and get some information out of it.

VI. IMPLEMENTATION

A. Gathering unstructured data

In this we have to extract tweets from the twitter the application . The possible keywords for searching the relevant tweets can be study, lab, homework. To extract the tweets from twitter we had to first create the twitter account. Then we are supposed to get the authorization to access the tweets. After getting the API

keys and token keys we can access the twitter tweets.

In our implementation we are using R programming language in RStudio. After getting the access to tokens and API's then in our tool we have to download the relevant packages for mining the tweets like twitteR, ROAuth, plyr, stringr, ggplot2. Then download the tweets and access them through OAuth package. Specify the filters or keywords on the basis on which you want to extract the tweets and save them to a csv file.

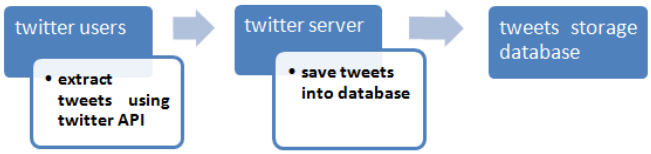

Fig.2. To illustrate the process of extracting tweets from twitter.

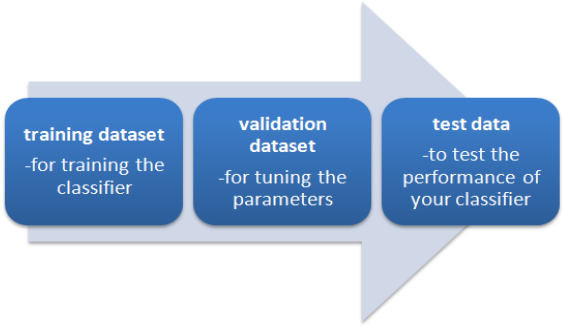

Fig.3. Datasets required for text mining application

B. Data cleaning

It is done to remove the inconsistencies and any kind of error in the textual data. We have to create a corpus and then remove the urls, words or characters other than English language, punctuations, stop words, extra whitespaces and numbers. All this operations are defined under tm package. Tm package is used for importing of the Data, handling of corpus, Creating and preprocessing of the document term matrix.

C. Splitting the data

The data is split into three parts as follows: Training dataset, validation and test dataset.

Figure 3 explains the use of different datasets that are used in text classification phase of the application.

D. Frequent words and association

We have to create a document term matrix and then inspect the frequent words in that matrix .Then we can plot the words with respect to their count as a graphical representation. Then we can find the words which are associated with each other.

E. Word cloud

It is made by counting the frequency of words and sort them in decreasing order of frequency. Then we can assign the colors to the words based the frequency count. Then the word cloud can be plotted. The package required for plotting it is wordcloud.

Figure 4 shown below on the next page represents the wordcloud based on the keyword "exams".

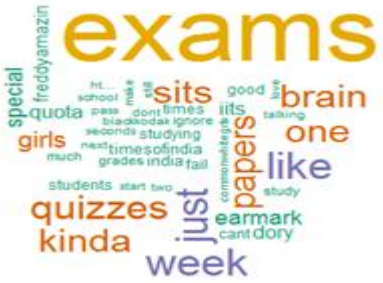

Fig.4.Wordcloud representation

F. Clustering

It is a technique to group the similar elements. This technique comes under the category of unsupervised learning in which there is no default topics or categories. Using this, the documents can appear under one or more subtopics , thus assure that the useful or important document is not deleted from the results after the search. A basic clustering algorithm will create a vector for the topics of each and every document and then assigning that document to the particular category under which it falls. Kmeans clustering algorithm can be use for the clustering of words into specific number of groups, such that the sum of squared distances lie between the respective word and the particular group. The number of groups you want to specify can be changed by changing the number mentioned in the kmeans() command. For clustering we need to install the cluster package .

G. Topic Modelling

In NLP and Machine learning, topic model is a statistical model for exploring the summarized topics which are there in the accumulated documents. It is an often used tool for text mining for getting the hidden semantics in the textual data. If we are given a document on a particular topic then we would predict that the word might occur more or less frequently in the documents. Topicmodelling package needs to be installed for this purpose and then we can apply the Latent Dirichlet Allocation.

H. Text Classification or Categorization

This is the last step for mining in which we can assign document to one or more class according to the content of the document. Classes are defined on the basis of the previously defined hierarchy of classes. Classification involves identifying the main themes of a document by placing the document into a predefined set of topics. It only counts words that appear and, from the counts, identifies the main topics that the document covers. There are various algorithms that are available for text classification like SLDA (Supervised Latent Dirichlet Allocation), SVM (Support Vector Machines), NaiveBayes algorithm, etc. In this paper we are using SVM Algorithm which is explained in the next section.

Here for classification in training set we had to provide labels manually to tweets and on the basis of that we have to test for the labels in the test data.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

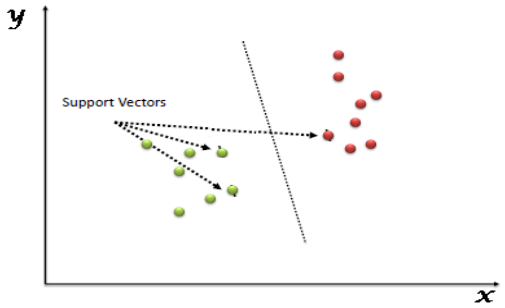

Fig.5. Plotting of the data items in particular coordinate according to feature and establishing the hyperplane to distinguish between two classes.

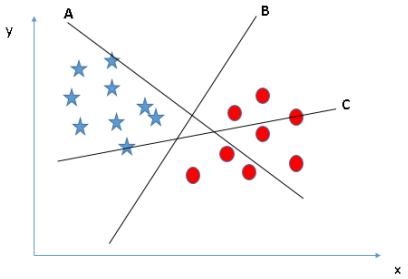

Fig.6.Segregating the classes on the basis of the best hyper plane.

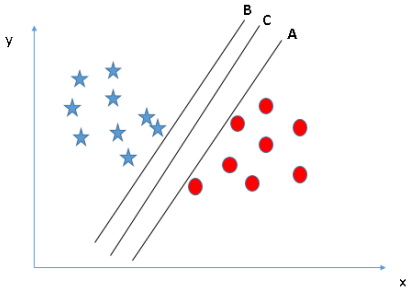

Fig.7. Finding the best hyper plane even if they are able to segregate the classes by finding margin distance.

I. Validation:

Using 5-fold cross validation to determine the accuracy of classifying the tweets on the basis of class labels. Here, Cross Validation means that we have to evaluate the predictive models by partitioning the data(original sample) into training data and test data so that we can evaluate it. In a K-fold cross validation we randomly sample the original data into k subsamples of equal size. Out of k sample one sample is used for validation test data and other k-1 samples are used for training data. For classification problems one uses ranked k-fold cross validation in which folds are selected such each fold will contain almost same proportion of class labels.

Accuracy = Total of true classification*100
Total of testing data

VII. SVM ALGORITHM

Support Vector Machines is an unsupervised learning algorithm by which both classification and regression can be done. In SVM, each data item is to be plotted as a point in an n dimensional space (where n is the number of features that are there for classification) where each feature is represented by a particular coordinate. Then the classification is performed by finding the hyper plane so that the hyper plane can differentiate the different type of classes well.

In Figure 5, we are establishing the hyperplane to distinguish between two classes and plotting dta items base on the features.

The plot shows the two different types of classes and how they are separated by the hyper plane. Here, the support vectors are the individual observations.

A. Identifying the Right Hyper Plane

- ***Case 1***

A thumb rule to identify the right hyper-plane: "Select the hyper-plane which segregates the two classes better".

In this case(Refer Figure 6), the hyper-plane "B" has performed this job very well.

- ***Case 2***

In this case we have three hyper-planes (A, B and C) and all are differentiated in the classes well.

In Figure 7, Finding the best hyper plane even if the classes are segregated by finding the margin distance.

The idea is to maximize the distance between the closest data point and hyper plane. This will help us to decide the appropriate hyper plane. This distance is termed as *Margin.*

Look at the figure below on the next page:

Here we can see that margin for hyperplane C is more as compared to both the planes A and B. Hence the right hyperplane is C plane. Another reason for selecting hyperplane with more margin is the robustness. If we select the hyperplane with low margin then there is a chance that the classification of the classes won't be correct.

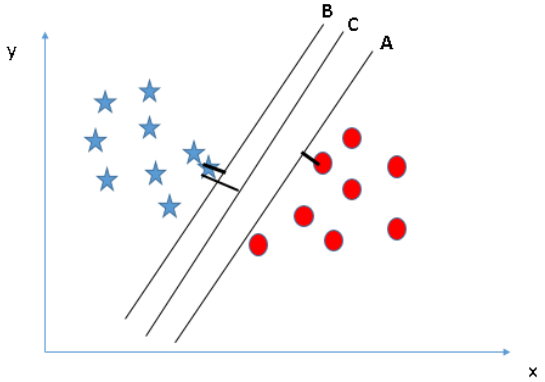

Fig.8. Choosing the hyper plane on the basis of maximum distance for classification.

B. SVM Using R

To create a SVR(support

vector regression model) we have to install the e1071 package and add the library(e1071) to the starting of your classification part. Support vector machines uses a function called svm for classifying text. If the data is categorical then function will automatically choose SVM. To improve the performance of the support vector model we need to select the best parameters of the model. This process of choosing the best parameters is known as Hyper parameter tuning or model selection. Grid search is the correct way of doing this. This means that we will have to train lots of models based on different parameters and choose the best out of it.

C. Why Should SVMs Work Well for Text Categorization?

• Input space is high dimensional: while learning text classification we have to deal with various number of features. This SVM has the capability of protecting by over fitting which hardly depends on the number of features.SVM can handle this huge number of feature spaces.

• Less number of extraneous features: if we assume that most of the data is irrelevant then we can avoid this these high dimensional input spaces. These things are done by feature selection. But in categorization of text we rarely get to see any unwanted features.

• Sparse Document vectors: few entries are not zero for the corresponding document vector di in each document di.

D. Pros associated with SVM

• Works well in high dimensional space.
• Does not consider extraneous features.
• Subset of training points are considered functions based on decisions which are known as support vectors. That is why it is efficient in terms of memory usage.

VIII. RANDOM FOREST

Random Forest is an adaptable, simple to utilize machine learning calculation that produces, even without hyper-parameter tuning, an awesome outcome more often than not. It is likewise a standout amongst the most utilized calculations, since it's straightforwardness and the way that it can be utilized for both arrangement and relapse errands.

It is a managed learning calculation. Like you would already be able to see from it's name, it makes a timberland and makes it some way or another irregular. The „forest" it constructs, is a group of Decision Trees, more often than not prepared with the "packing" technique. The general thought of the stowing strategy is that a mix of learning models expands the general outcome.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. I will talk about random forest in classification, since classification is sometimes considered the building block of machine learning.

With a couple of exemptions an irregular timberland classifier has all the hyperparameters of a choice tree classifier and furthermore all the hyperparameters of a sacking classifier, to control the group itself. Rather than building a packing classifier and passing it into a choice tree-classifier, you can simply utilize the arbitrary timberland classifier class, which is more helpful and enhanced for choice trees.

The arbitrary woodland calculation carries additional irregularity into the model, when it is developing the trees. Rather than hunting down the best component while part a hub, it scans for the best element among an irregular subset of highlights. This procedure makes a wide decent variety, which for the most part brings about a superior model.

Accordingly when you are growing a tree in arbitrary woods, just an irregular subset of the highlights is considered for part a hub. You can even make trees more arbitrary, by utilizing irregular edges over it, for each component as opposed to hunting down the most ideal edges (like a typical choice tree does).

A. Advantages and disadvantages:

favorable position of irregular timberland is that it can be utilized for both relapse and order undertakings and that it's anything but difficult to see the relative significance it allots to the information highlights.

Arbitrary Forest is likewise considered as an extremely convenient and simple to utilize calculation, since it's default hyperparameters regularly create a decent expectation result. The quantity of hyperparameters is likewise not that high and they are clear to get it.

One of the huge issues in machine learning is overfitting, yet more often than not this won't occur that simple to an arbitrary timberland classifier. That is on account of if there are sufficient trees in the woodland, the classifier won't overfit the model.

Table 1: Accuracy of classification algorithms

S.No.	Algorithm	Accuracy
1.	SVM	82.2%
2.	Random Forest	93.3%

The fundamental restriction of Random Forest is that countless can influence the calculation to ease back to and inadequate for constant forecasts. All in all, these calculations are quick to prepare, yet very ease back to make forecasts once they are prepared. A more exact expectation requires more trees, which brings about a slower display. In most true applications the arbitrary woodland calculation is sufficiently quick, however there can positively be circumstances where run-time execution is vital and different methodologies would be favored.

Furthermore, obviously Random Forest is a prescient displaying device and not a spellbinding instrument. That implies, on the off chance that you are searching for a portrayal of the connections in your information, different

methodologies would be favored.

B. Random Forest Creation Pseudo code

- Randomly select "**K**" features from total "**m**" features where **k << m.**
- Among the "**K**" features, calculate the node "**d**" using the best split point.
- Split the node into **daughter nodes** using the **best split**.
- Repeat the **a** to **c** steps until "**l**" number of nodes has reached.
- Build forest by repeating steps **a** to d for "**n**" number times to create "**n**" number of trees

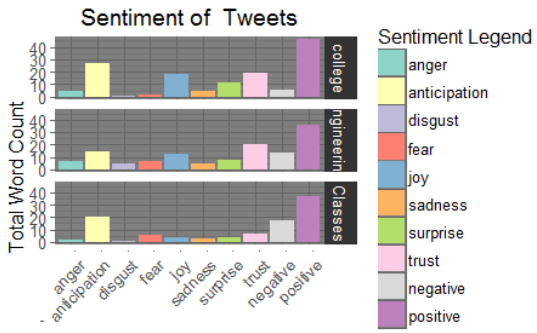

Fig.9. Sentiment Analysis of tweets based on the emotions.

IX. EXPERIMENTAL RESULTS

Here we are showing the results of sentiment analysis in two ways:

1. Showing the count of tweets under the categories of different types of emotions.
   Figure 9 shows the count of tweets falling under different categories of emotions.
2. Variation in the tweets over the period of the time.
   Keywords used for this are: college, engineering, classes.
   Figure 10 shows the variation in the sentiments of tweets over a period of the time.
3. Percentage of Accuracy of the classification Algorithm.
   Here our main focus was on getting the results using SVM Algorithm but after comparing the results with ensemble learning method i.e. Random Forest we found that it has better results.(Refer Table 1).

X. CONCLUSION

This paper conveys about what is text mining and how it is done. It consists of basic outline of all the steps of the text mining. We have also given an outline about analyzing the problems of students in their learning experiences which can inform educationists and others makers to gain further understanding of the students experiences in their college life.

We have also tried to classify the tweets in test data based on the labels that we have in training data and predicted the accuracy of the supervised learning paradigms.

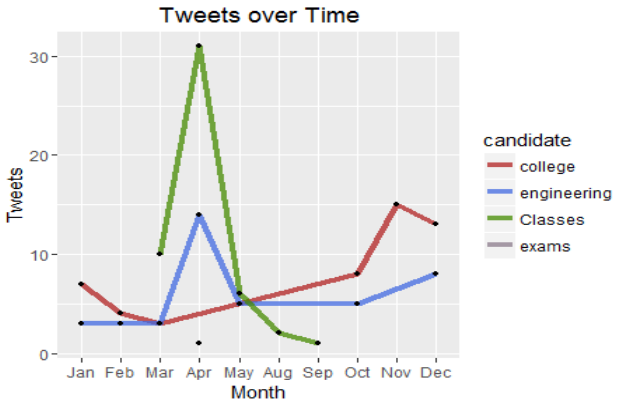

Fig.10. Variation in the sentiments of tweets over the period of time

REFERENCES

1. Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences ",JULY-SEPTEMBER2014
2. Antonio Moreno1, Teófilo Redondo2 1Universidad Autónoma de Madrid and Instituto de Ingeniería del Conocimiento, Madrid, Spain 2ZED Worldwide, Madrid," Text Analytics: the convergence of Big Data and Artificial Intelligence",Spain-2016
3. Florian Heimerl, Steffen Lohmann, Simon Lange, Thomas Ertl Institute for Visualization and Interactive Systems (VIS) University of Stuttgart, "Word Cloud Explorer: Text Analytics based on Word Clouds",Germany 2014
4. Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse G¨oker, Yiannis Kompatsiaris, Alejandro Jaimes," Sensing trending topics in Twitter"
5. S. Buckley M. Ettl P. Jain R. Luss M. Petrik R. K. Ravi C. Venkatramani, "Social media and customer behavior analytics for personalized customer engagement"
6. Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, Senior Member, IEEE, and Shojiro Nishio, Fellow, IEEE," "Wikipedia-based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes"
7. Jie Yin, CSIRO ICT Centre, Andrew Lampert, Palantir Technologies, Mark Cameron, Bella Robinson, and Robert Power, CSIRO ICT Centre "Using Social Media to Enhance Emergency Situation Awareness"
8. Donald E. Brown, Fellow, IEEE. "Text Mining the Contributors to Rail Accidents"
9. Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, Member, IEEE, and Francesco Marcelloni, Member, IEEE, "Real-Time Detection of Traffic From Twitter Stream Analysis"
10. Yuya shibuya, graduate school of interdisciplinary information studies, "mining social media for disaster management: leveraging social media data for community recovery"
11. E Rejeesh , M Anupama , Computer science mahatama gandhi colege – iritty , kerala, india ,"Social media and data mining enabled pre-counseling session: A system to perk up effectiveness of counseling in distance education"