IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol.6, No4, July-August 2016

ANALYSIS OF STUDENT BEHAVIOUR USING R LANGUAGE

Ashish Soni
Asst.Professor
Malwa Institute of Science & Technology
Indore,India
ashishsn9@gmail.com

Toshi Sharma
MTech Scholar
Malwa Institute of Science & Technology
Indore,India
toshishm.25@gmail.com

Vishal Chhabra
Malwa Institute of Science & Technology
Indore,India
vishalchhabra26@gmail.com

Abstract: **Today Social Media have become the greatest fascination among the youth. The amount of content generated on these sites (Twitter, Facebook, Linkedin) is not only enormous but also relevant from research point of view. Students post about their feelings on these social platforms fearlessly. The informal conversation outside the controlled classroom environment could be of great help for educational researchers. Based on these we drew out a sentiment analysis of engineering student problems using R as a platform.**

Keywords: Big Data, R language, Twitter, #EngineeringProblems, #EngineeringStudents

I. INTRODUCTION

Social media have become a biggest platform where students express their joys, sorrows, anger informally without the fear of their teachers, parents or peers. Their tweets, posts, comment provide a greater insight into what their present mental state is. Thus this data can prove to be of great help to educational researchers to identify at risk students and their problems. This in turn can help to formulate policies based on individuals to help improve learning process. Pure manual analysis cannot deal with the ever growing scale of data, while pure automatic algorithm cannot capture in-depth meaning within the data [1].

Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, classroom activities to collect data related to students learning experiences [2][3]. Thus cannot prove that much beneficial with the growing student population. Also the reliability of such experiment is a great issue for concern, as the students may not be honest in the above mentioned analysis. But when it comes to expressing their beliefs on social media, students are found to be more open than in the controlled classroom environment.

The emerging field of learning analytics and educational data mining has focused on analyzing structured data obtained from course management systems (CMS), classroom technology usage, or controlled online learning environments to informal educational decision making. However as per our research there is no study done to analyze or classify the bad and good posts or tweets to help take decision in education using data mining.

The main aim of our study is to obtain a clear understanding of how social media data can help improve the learning process using data mining techniques and analysis. Our main focus was on engineering students who active on Twitter because:

1. Engineering has been described as the toughest course with several assignments, tests, papers and projects.
2. There has been a significant drop in the no of students pursuing engineering in recent years.
3. The reason for choosing Twitter is that many engineering students are active on Twitter and posts there feeling

regularly. Also of all other sites available it is considerably easy to mine data on Twitter (free API's).

Architecture workflow:

The data obtained is raw data. Therefore some text pre-processing is required on the data, so that we can obtain some useful data from it. For this R provides "tm" package which can be used to process the data. It has a function called tm_map() to process the data.

The architecture workflow is as follows:

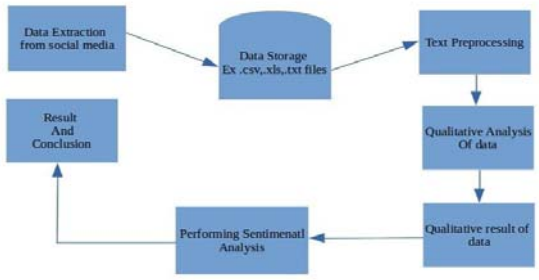

Figure 1: Architectural Workflow Diagram

After processing we can now move on to our final sentiment analysis. Sentiment analysis is the process of classifying tweets (positive or negative). It totally analyses all the words present in each individual tweets and classify them as positive or negative. Then according to the frequency of positive or negative words it categories the tweet as positive or negative whichever words are present in a greater quantity.

II. RELATED WORK

A. Getting The Data From Twitter

To obtain the data from twitter we first created an account on Twitter. Then using the developer's option of twitter, we build an app on twitter. After the app is created Twitter provides Access Key, Access Token, Consumer Key, and Consumer Secret which is used to uniquely identify an entity. Using these credentials you need to authenticate yourself using the Twitter OAuth. After we are done with the authentication, we can now use simple methods to retrieve the data. We have used two hash tags #engineering problems and #engineering students. The reason for choosing these hash tags is that these are the most important hash tags if we want to know the conditions of engineering students.

B. Perform Preprocessing

The data obtained is unfiltered and contains many irrelevant things like urls, numbers, stop words etc. that needs to be removed before processing the actual data that needs to be removed before the actual processing of data can be performed.

For this R defined a variety of pre-defined packages with several methods to perform processing.

C. Sentiment Analysis

Sentiment Analysis is the process of classifying tweets whether positive or negative based upon the frequencies of the words appearing in the tweets. It is one of the most important features provided by the R language. It can be a great help for estimating the response, views of the customers towards a particular brands or products. The sentiment analysis uses Navies Bayes Algorithm.

Navies Bayes Algorithm:

Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, X = {x1,x2,x3...,xd}, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1,c2,c...,cd}. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

p(Cj|x1,x2,.....,xd) α p(x1,x2,.....,xd | Cj) p(Cj)

where p(Cj|x1,x2,x3...,xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj.

Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood to a product of terms:

p(X|Cj) α dΠk=1 p(xk| Cj)

and rewrite the posterior as:

p(Cj|X) p(Cj) α dΠk=1 p(xk| Cj)

Using Bayes' rule above, we label a new case X with a class level Cj that achieves the highest posterior probability.

D. Analysis

The analysis was first done on the sample data consisting of some 189 tweets. This was done to check the relevancy of the system. Also smaller set of data offer greater accuracy.

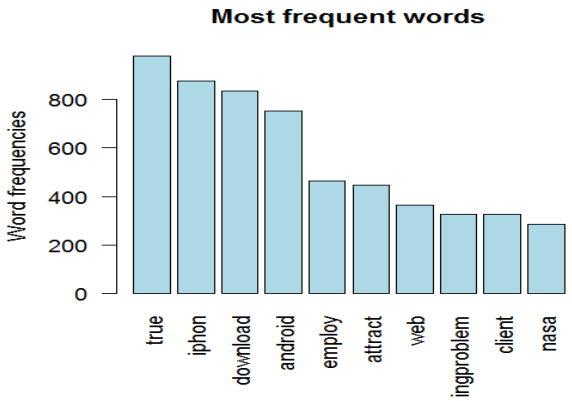

Figure 2: Sample Histogram

Figure 3: Sample Word Cloud

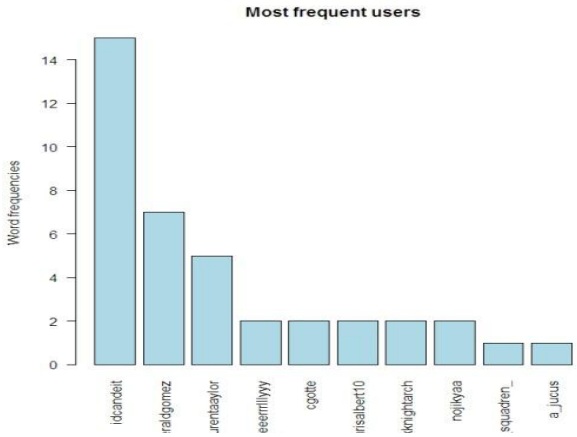

Figure 4: Sample graph for most frequent users

After that we took 1933 tweets from the #engineeringproblem and #enginerringstudent and obtained wordcloud, most frequent words appearing in the tweets, and the most affected users. The visual representation of the final analysis is as follows:

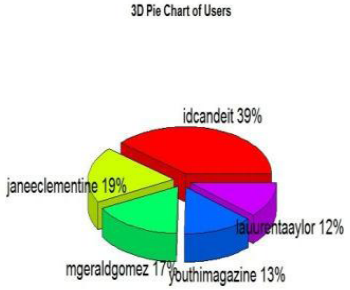

Figure 5: Affected Users Graph

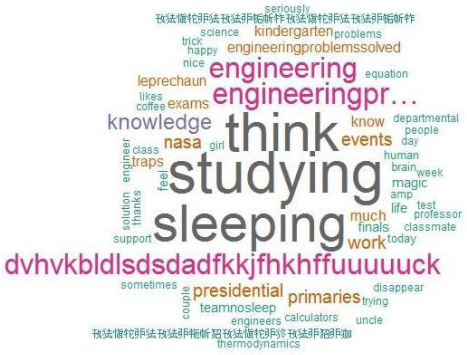

Figure 6: Actual Word Cloud

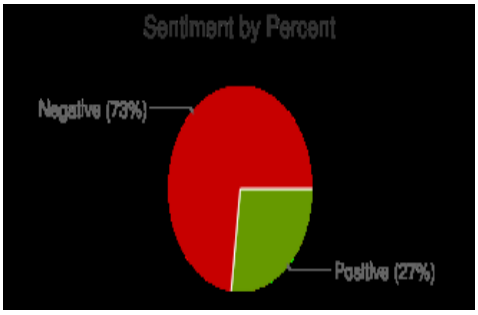

Figure 7: Sentiment Categorization of Students

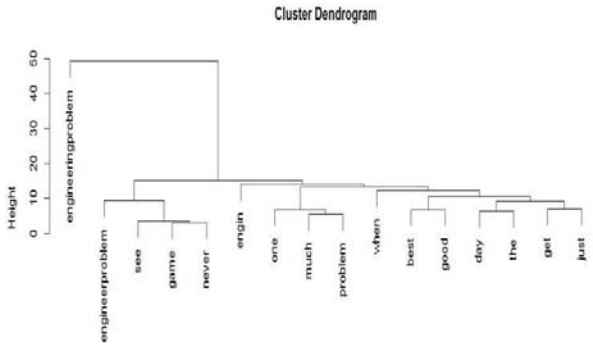

Figure 8: Cluster Dendrogram

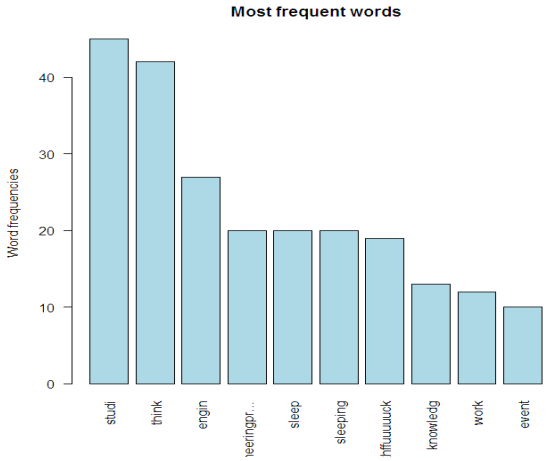

Figure 9: Most Frequent Words

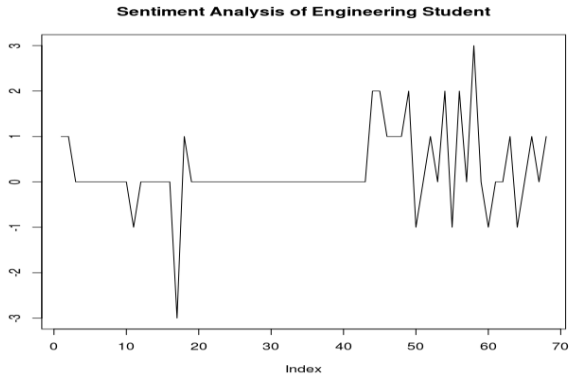

Figure 10: Sentiment Analysis of Students

III. RESULT OF THE ANALYSIS

The analysis on the tweets clearly points the most affected users. These are the users who need special attention and care. The sentiment analysis on the tweets show that around 73% of the tweets were found negative and only 27% were positive. Thus we can conclude that engineering students are really pressurized. Analysis of the word cloud shows that, the major problems of the students are related with the sleep. Various assignments, project pressure hamper the sleep of the students. E.g. of some tweets found on sleep problems are:

1. RT @HugotInhinyero: I think of sleeping while studying
2. RT @engineerproblem: Looking forward to going to sleep is the best part of my day. #engineeringproblems
3. RT @Lady_Engineers: Up all night to get lucky. (Hopefully, on this test tomorrow) #engineeringproblems #ladyengineer

Other problems that students face are that of peer pressure. People's expectation about engineers is what is reflected in most tweets. For e.g. I don't understand, just because I'm an engineering student, people associate with me that I know EVERYTHING, which is not true. Besides these problems regarding exams, studies etc have also been highlighted in many of the tweets as highlighted by the word cloud.

IV. DISCUSSION, LIMITATIONS AND FUTURE WORK

Our studies reveal the use of social media data with big data analytics to find the behavior of students in engineering and take necessary actions. Through our study we came to the conclusion that engineering students face a problem of heavy study load and high completion. Therefore, they are not able to sleep properly. Because of lack of sleep there might be issues with students like health problems, low retention, lack of concentration and even depression.

This is the reason for high rate of suicides among engineering students. A detailed analysis could not be done in case of India because not many students are active on Twitter probably, because Twitter is relatively complex to understand and mostly considered to be used by people having high intellect and famous personalities in case of India. Thus Facebook and other popular sites can be taken into considerations for future work.

Also data of specific institutions can also be used to find the problems that students face in specific institutions. Future work may also take into considerations students belonging to different backgrounds (like law, management, medical etc.) and present a comparative analysis. This will help to identify the most pressurizing course. Few of the limitations of our

work include lack of awareness among students in managing their identity on social media.

V. CONCLUSION

Our study might prove beneficial to educational researchers and institutions suffering from decline in the ratio of students taking admission in engineering. It provides a fast and accurate to analyse the behaviour of students. The no of human efforts required are also reduced considerably.

The data obtained is raw data. Therefore some text pre-processing is required on the data, so that we can obtain some useful data from it. For this R provides “tm” package which can be used to process the data. It has a function called tm_map() to process the data.

VI. ACKNOWLEDGEMENT

With due respect we would like to thank to our advisor Prof. Vishal Chhabra for his support and guidance because of which we are able to propose this article. His thoughtful questions and comments were valued greatly. It would not have been completed without the encouragement of Prof. Chhabra.

VII. REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu and Krishna Madhavan, “Mining Social Media Data for Understanding Students’ Learning Experiences,” IEEE, Volume 7, Issue 3, 2014.

[2] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Streveler, and K. Smith, “Academic pathways study: Processes and realities,” in Proceedings of the American Society for Engineering Education Annual Conference and Exposition,2008.[1] Xin Chen, Mihaela Vorvoreanu and Krishna Madhavan, “Mining Social Media Data for Understanding Students’ Learning Experiences,” IEEE, Volume 7, Issue 3, 2014

[3] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Streveler, and K. Smith, “Academic pathways study: Processes and realities,” in Proceedings of the American Society for Engineering Education Annual Conference and Exposition,2008.

[4] C.J. Atman, S.D. Sheppard, J. Turns, R.S. Adams, L. Fleming, R. Stevens, R.A. Streveler, K. Smith, R. Miller, L. Leifer, K. Yasuhara, and D. Lund, “Enabling engineering student success: The final report for the Center Advnacement of Engineering Education,” Morgan & Claypool Publishers, Center for the Advancement of Engineering Education, 2010.

[5] M.Vasuki, J.Arthi, K.Kayalvizhi, “Decision Making Using Sentiment Analysis from Twitter, ”International Journal of Innovative Research in Computer and Communication Engineering, Vol 2, Issue 12, December 2012.