Elicitation of Student Learning Experiences from Twitter using Data Mining Techniques

Pratiba D[1], Samrudh J[2], Dadapeer[3], Srikanth J[4]

Asst. Professor, Dept. of Computer Science & Engineering, R V College of Engineering, VTU, Bangalore, India[1]
Student, R V College of Engineering, VTU, Bangalore, India [2,3,4]

Abstract - *Big Data analytics facilitates entities to analyze a combination of data which may be structured, semi structured or unstructured. It acts as an added value to business related information and also provides an insight about various aspects of the respective business. Big data Analytics needs excellent technology to resourcefully process huge amount of data within acceptable elapsed times. The informal conversations carried out by students on various Social Media platforms such as Facebook, Twitter etc give a lot of information on their experience of education along with a major highlight on their opinion and concerns about the system of learning. Data from these kinds of platform and environment give an unbiased view and helps in improving the experience of student learning. But, the real challenge lies in analyzing and drawing conclusion from this kind of data. Here, human interpretation of these student experiences becomes a requirement. But, the ever increasing data stresses on the need to have automatic techniques for analysis of the data. The proposed system develops a workflow to put together both, data mining techniques at a large scale and an adequate amount of qualitative analysis. This paper puts the spotlight on posts by engineering students on Twitter and its analysis to comprehend problems in their educational experiences.*

Keywords - *Big Data, Social Media, Data Analysis, Student Learning Experiences*

I. Introduction

Various Social media platforms like Twitter and Facebook act as a place for students to share their views and emotions. Students discuss about everything in a casual and informal way on these kind of sites. These digital footprints of students become the source for a large amount of implicit knowledge. It provides a different and new perspective for educational researchers and practitioners to understand the experiences of students outside the classroom atmosphere [1].

This knowledge helps in decision-making with respect to improving quality of education, thereby ensuring better student recruitment along with a higher rate of retention and success. The huge amount of data from social media along with giving a chance to understand experiences of students brings in methodological difficulties in getting only the data for educational purposes.

The various issues include use of slangs, the unpredictability of place and time of the posts, understanding the students mindset etc. Just manual analysis cannot deal with the ever growing data and at the same time, purely automatic algorithms generally cannot capture in-depth meaning within the data. The upcoming area of analytics and educational data mining has laid importance on analyzing structured data attained through Course Management Systems, usage pattern of classroom technology, online learning.

However, there is no research particularly for mining or analyzing content posted by students specifically to study and understand students' learning experiences. The goal of this study include demonstration of workflow of social media data integrating data mining techniques and qualitative analysis. It also aims at exploring informal conversations of engineering students and throwing light on problems faced by them in the learning process. The reason behind choosing engineering students and their posts is that engineering schools have been facing problems in student recruitment and retention [3]. Also, Engineers play a significant role in the country's growth and also have an impact on economic growth.

II. RELATED WORK

In the last decade, online social networks (OSNs) have become very popular worldwide. The E-Learning techniques have made the process of learning convenient and reachable through these networks. Nevertheless, combining OSNs with E-learning is a new idea. The role of OSNs in E-Learning experience is dealt with[1].

For businesses, spreading of word online has become an important resource. Various aspects include analyzing user generated reviews, classifying them into sentiment classes etc and this is something everyone has started paying attention to. At present, there is some research on sentiment analysis for English traveler generated reviews but there is very little work and research on reviews in other languages. [4]

Facebook contains enormous data of almost 5 billion pieces of them shared by over 400 million users every month. Analyzing this huge amount of unstructured data brings a lot of challenges. For example, consider GraphCT, a Graph Characterization Toolkit for massive graphs representing social network data. On a 128-bit processor Cray XMT, GraphCT approximately processes a real world graph with 61.6 million vertices and 1.47 billion edges in about 100 minutes [5].

Analyzing online data from social networks provides opportunities for extracting attributes of sentimental influence, It uses models to study sentimental influencing as well as influenced probabilities for users of the popular online social media, Twitter. It is found that there is a high correlation between influencing probabilities and influenced probabilities [6].

The drawbacks of the surveyed papers are: Privacy and security of personal information, lack of multimedia content support, grouping of large volume of data, deals only with static influence computing and the algorithms such as KNN, simple logistic regression and SVM takes more time to classify.

III. Problem Statement

In today's world, a lot of focus is on the study habits and study processes which improve the knowledge of the students. But, there are no approaches which mainly concentrate on the mental health of students. Students have to be happy and enjoy their life along with concentrating on the career prospects. Also, with the advent of social media applications like facebook, twitter have a lot of sentiments placed by various people at different age levels.

This huge amount of data can be very useful for various conclusions ranging from something as primitive as the most liked actor. In this project, students' learning experiences i.e. the sentiments are taken from twitter and then analyzed to make conclusions about the problems faced by students and take steps to resolve them by using '*Naïve Bayes Data Mining Algorithm*'.

The system is used to show the workflow of social media data which is useful for educational purposes. This is done by integrating both qualitative analysis and large-scale data mining techniques to explore informal conversations of students on Twitter. [2].

IV. Definitions

A. *Student Behaviour*

According to Wikipedia, student behaviour means combination of every physical action and observable emotion of a student [7]. Some traits change with age whereas some of them which are specific to an individual's personality will be consistent. But, behaviour which is driven by thoughts and feelings gives an insight into a person's psychology. Social behaviour of a student helps to study the influence of social interaction, ethics and culture on the student.

B. *Twitter as Social Media*

Twitter is the one of the most important social media where lot of discussions, conversations and comments are made. Now-a-days many students are using twitter to make comments on the education system. Some comments are for and some against. Hence, to understand student behavior these comments are analyzed in this system.

V. CHALLENGES

- Thoughts and feelings of only the active students on Social Media or Twitter to be specific gets logged.

- The mined data requires further scrutiny as it doesn't only contain problems and complaints faced by the Engineering students but also some good things about engineering under hash tags such as '#EngineeringPerks' [4].

- Only relatively prominent themes with higher tweets are identified. There may be a variety of other issues hidden which are of importance to the researchers.

- Existence of correlation among the themes is observed by qualitative analysis. For example, 'heavy study load' can lead to 'lesser social engagement' and 'sleep problems.' Similarly, 'negative emotion' may be caused by some other themes. Naïve Bayes classifier is built by label independence assumption. The classifier used is designed to be a multi-label classifier in order to reconcile this effect. The comparison experiment with M3L shows that the advanced model that accounts for label correlations does not perform as well as the simple Naïve Bayes model [8].

VI. Existing System

Educational researchers have been using methods such as focus group studies, interviews and surveys to collect data related to students' learning experiences. These methods are time consuming and cannot be done very frequently. Also, the scale of such studies is limited. Further, when prompted about their experiences, students reflect on what they were thinking and doing in the past which may have become obscured over time[2].

VII. Proposed System

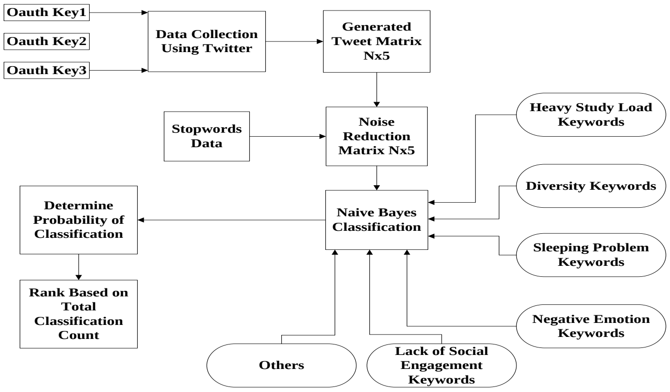
Oauth Key1
Oauth Key2
Oauth Key3
Data Collection Using Twitter
Generated Tweet Matrix Nx5
Heavy Study Load Keywords
Stopwords Data
Noise Reduction Matrix Nx5
Diversity Keywords
Determine Probability of Classification
Naive Bayes Classification
Sleeping Problem Keywords
Rank Based on Total Classification Count
Negative Emotion Keywords
Others
Lack of Social Engagement Keywords

Fig. 1 : System Architecture of Elicitation of Student's Learning Experiences from social media data

The above figure represents the architectural diagram of Elicitation of Students' Learning Experiences System. The following sections explain each module of the system architecture.

A. *Data Collection using Tweet*

The sentiment/tweets are collected from a set of 20 accounts. The data retrieval is done using twitter API, OAuth api used to authenticate the open source with the twitter application.

B. *Sentiment Storage based on Tweets*

The sentiment storage based on Tweets is a process of storing the data about the tweets into the relational storage in terms of *(Twitter Id, Twitter Desc, and User Id).* Twitter Id is unique Id associated with the tweet, TwitterDesc is the actual tweet and UserId is the Id associated with the user.

C. *Stopwords*

These are the set of words which do not have any specific meaning. The data mining forum has defined a set of keywords for this and they are filtered out while processing the data.

D. *Data Cleaning*

Data Cleaning is used for removing the stop words from each of the tweets and clean them. After the data cleaning process is completed the clean data can be represented as a set *(Clean Id, Clean Data, and User Id). Clean Id* is the unique Id associated with the Tweet, *Clean Data* is the clean data after removal of clean data and *UserId* is the unique Id associated with the user.

E. *Maximum Likelihood Method*

For each of the tweets under each of the category the probability is computed using the following rules [9]. An example is as below:

1. The number of categories is 5. The categories can be a set say, C= {c1, c2, c3, c4, c5}
   where c1 = heavy study load
   c2 = lack of social engagement
   c3 = negative emotion
   c4 = sleep problems
   c5 = diversity issues

2. The number of words present in each tweet is denoted as N.

3. For each of the words the two probabilities are computed using
   a) The probability of this word in a specific category *c* is calculated using pre defined formula i.e. p (c /d).
   b) The probability of this word in categories other than *c* is given by p (!c /d).

4. Define a specific threshold called T.

5. If the probability is greater than T, then the word belongs to the specific category. For all categories the probability is computed and then the word is labeled to the category it belongs to T.

6. The contingency table is created for each of the words.

7. The accuracy for the word is computed by using the formula below

$$accuracy = \frac{t_p + t_n}{t_p + t_n + fp + fn'}$$

8. The Recall, r and Precision, p of the word is measured by using the formula below

$$r = \frac{t_p}{t_p + fn'} \qquad p = \frac{t_p}{t_p + fp'}$$

9. The F1 measure is computed using

$$F1 = \frac{2.p.r}{p + r} = \frac{2t_p}{2t_p + fp + fn'}$$

10. This is then followed by Micro and Macro averaging after which the tweets are classified into one of the 5 categories.

VIII. Algorithms

A. *Chi-Squared Algorithm*

In probability theory, the chi-squared distribution with n degrees of freedom is defined as the distribution of the sum of squares of n independent standard normal random variables. It is a largely used distribution in inferential statistics such as in constructing confidence intervals and hypothesis testing. It is a specific case under gamma distribution[10].

The above said distribution finds application in chi-squared tests for goodness of fit of an observed distribution to a theoretical one. Many statistical tests like Friedman's analysis of variance by ranks also use this distribution.

This algorithm is mainly used to Reduce Noise from the tweet which is nothing but removing the stopwords as it is very important to reduce the time complexity. There are around 1013 words defined as stopwords [11].

B. *Naïve Bayes Algorithm*

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, shown as feature valued vectors. It requires not one but many algorithms to train the classifier. It is based on a common principle: all naive Bayes classifiers assume that the value of a certain feature is not dependent on the value of any other feature, given the class variable. For example, a fruit may be considered to be an orange if it is orange in colour, round, and about 5cm radius. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an orange, regardless of any possible correlations between the color, roundness and radius features [12].

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

The above formula is used as a probability model to calculate the probability of words belonging to particular category.

Finally, the probability of category based on the data given in each tweets is computed using above formula.

Then based on defined frequency classification is made. Even three parameters like contingency, Enhance contingency and its count are defined.

IX. Results and Anaysis

This section lists the result and the inferences made from the testing results. The evaluation metrics have been listed and the results have been accordingly quantified. Several data sets were taken under consideration and models were built and analyzed.

A. *Evaluation Metric*

The objective of the Student Learning Experience is to collect the real time tweet data and analyze the same such that which tweets belongs to what category. Hence the important evaluation metric for this project is the time taken to classify the tweet data to specific category. They are four evaluation metrics as follows [13].

- **Accuracy**: It is defined as the ratio of the no. of accurate cases to the total no. of cases.
- **Precision**: It is the fraction of retrieved instances which are correct and relevant.
- **Recall**: It is the fraction of relevant instances that are retrieved or the percentage of correct items that are selected.
- **F Measure**: A metric that combines precision and recall metrics. It is the weighted harmonic mean.

B. *Experimental Dataset*

The discussed implementation was applied on a data set over a defined period of days. The data set consists of thousands of tweets. Each tweet data identified with unique ID, description and screen name of the particular tweet data. The input parameter list for the operation was given through the graphical user interface that was executing at the machine. The dataset file is also used for comparative analysis [14].

C. *Performance Analysis*

The testing results of trained models have been tabulated. The training accuracy and testing accuracy have been noted along with the 'c' and 'g' values

[15].True positive rate and True negative rate have been calculated. Also the number of descriptors used has been specified. The following table shows the performance of classifiers.

Table I Performance analysis of different classifiers

Classifiers	Accuracy (%)	Classification Time (Seconds)
Naïve Bayes	81.7	0.805
K-NN	67.55	2.198
SVM	69.6	2.248
Simple Logistic	98.75	45.368

X. CONCLUSION AND FUTURE WORK

This study is beneficial to researchers in learning analytics, educational data mining and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user generated textual content. The study can inform educational administrators, practitioners and other relevant stake holders to gain an understanding of college experiences of engineering students [16].

The system which initially aimed at making use of uncontrolled social media space, proposes some expected directions of future work for researchers. At the same time, it advocates that major attention should be given for protection of students' privacy while trying to provide good education and services to them. The 'manipulation' of personal image online needs to be taken into consideration. Future work can give a clear picture about both, the positive and negative aspects to investigate the tradeoffs which students struggle with.

Provided that the students complain about issues on social media, it may act as the platform for seeking support. Therefore, future work can be done on why and how students seek social support on social media sites too. Another aspect could be to address the correlations among the student problems. This could be a research direction where algorithms can be designed to set up the correlation. These will ultimately help the stakeholders of the education domain. It will contribute to the nation's growth positively.

REFERENCES

[1] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *Educause Review*, vol. 46, no. 5, pp. 30–32, 2011.

[2] Cheng mingzhi, Xin Yang, bao Jingbing, Wang Cong, Yang Yixian,: "A Random Walk Method for Sentiment Classification", proceedings of Second International Conference on Future Information technology and management Engineering,2013 IEEE conference, Sanya, Dec 13-14,2013,pp.327-330

[3] David Ediger, Karl Jiang,Jason Riedly, David A Bader, Courtney Corley Rob Farber William N Reynolds,"Masisve Social Network Analysis: "Mining Twitter for Social Good", IEEE 39th International Conference on Parrel Processing, San Dego CA, Sep 13-16,2013,pp.583-593

[4] 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, pp. 357–362.

[5] 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.

[6] 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. Ya suhara, and D. Lund, "Enabling engineering student success: The final report for the Center for the Advancement of Engineering Education," Morgan & Claypool Publishers, Center for the Advancement of Engineering Education, 2010.

[7] Loo Hanley, Timothy Ong Chee Aik, Raymond Wee Keat Kheng & Lim See Yew, "Mining Twitter Data to understand student behavior" IEEE 63rd Annual Conference International Council for Educational Media, Myanmar, c 5-8.2011,125-223

[8] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," *Knowledge Media Institute, Technical ReportKMI-2012-01*, 2012.

[9] Pallavi K., Pagare Department of Computer Engineering, MET's Institute of Engineering, Nashik, Savitribai Phule Pune University, Maharashtra, India, "*Analyzing Social Media Data for Understanding Student's Problem International Journal of Computer Applications*"(0975 –8887) Innovations and Trends in Computer and Communication Engineering (ITCCE-2014)

[10] Huang Sui, You Jianpinh, Zhang Hongxian, Zhou Wei, "Sentiment Analysis of Chinese Micro-blog using Semantic Sentiment Space Model", Proceeding of 2nd International Conference on Computer Science and network technology Guangzhou, China, Jul 12-26.2012, Vol. 1,pp.512-614

[11] Seyed-Alii, Bahrainian, Andreas Dengael, " sentiment Analysis using Sentiment features", Proceeding of International Conference on Web Intelligence(WI) and Intelligent Agent Technology(IAT), Germany, Aug 18-25.2012,pp.1040-1050

[12] Rabia Batool Asad Masood Khattak, Jahanzeb maqbool and Sungyoung Lee, kung Hee," Precise Tweet Classification and Sentimantal Analysis", Procedding of International Joint Conference on computer Science, South Korea, Dec 12-15.2012,pp.1204-1236

[13] Beiming Sun, Vincent TY, "AnalysingSentimental Influence of Posts on Social Netwroks",Proceedings of IEEE 18th International Conference on Computer supported Cooperative Work, Cairo, Egypt, Sept 29-30,2010.pp,23-24

[14] Ana Mihanovic,Hrvoje Gabelica, Zagreb, Croatia, "Big Data and Sentiment Analysis using KNIME: Online Reviews vs Social Media " ,Proceedingsof International Conference, Croatia, Mar 30-31,2010.pp.345-360

[15] F. Santos-Sanchez, Member, IEEE and A. Mendez-Vazquez .Members, IEEE, "Sentimental Analysis for e-services" IIAI 3rd International Conference on Advanced Applied Informatics , Kigali, Uganda, Jan 12-13,2010.pp.1120-1134

[16] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.