MINING SOCIAL MEDIA DATA FOR UNDERSTANDING STUDENTS’ LEARNING EXPERIENCE

Parth Harkhani
Dept. of Information Technology
Vidyavardhini’s College of Engineering and Technology, VCET
University of Mumbai
INDIA
Email id: harkhaniparth@gmail.com

Amol Duraphe
Dept. of Information Technology
Vidyavardhini’s college of engineering and technology, VCET
University of Mumbai
INDIA
Email id: amol.duraphe786@gmail.com

Prathamesh Kadam
Dept. of Information Technology
Vidyavardhini’s college of engineering and technology, VCET
University of Mumbai
INDIA
Email id: pratsstats9512@gmail.com

Anagha Patil
Dept. of Information Technology
Vidyavardhini’s college of engineering and technology, VCET
University of Mumbai
INDIA
Email id: anagha.patil83@gmail.com

ABSTRACT:- Nowadays students’ informally have their conversation through social media (example: Twitter & Facebook) about any topic right from interesting movies, news and to the studies and their learning and this shed light on their education related experiences—feelings, opinions and concerns about the learning process. The workflow of paper has been developed by us in order to integrate data analysis and data mining techniques for large amount of data. Now, students have a habit of making short frequent posts and for this Twitter is the largest and the best platform for the students to share their views about learning because it allows for only 140 characters, thus we have a tendency to provide example as engineering students’ Twitter posts to grasp drawback occurred in their academic expertise and also the problems they face during their learning. In this project we will be performing analysis on the tweets that we will collect of engineering students and know what problems they face in engineering college life. On the basis of these results, we will apply naïve bayes multi-label classifier so that tweet can classify into more than one category and tweets will be categorize that shows the students’ problems and focus on which category is having the largest number. This work basically provides us with a lot of information about the students and what their views and opinions are towards the education system and also what are the changes that are to be made through their informal conversation and this in-turn provides an insight into their learning experience.

I. INTRODUCTION

Data mining studies have effectively created many different methods, various ways for supervising large amount of data in response to difficulties faced. So data we fetch from this sites can be used for data mining[1]. This Students’ tweets provides high amount of data that shows how students faces the problems in education system and result can be given to education administrator so that they can take proper decision on it and they can improve the quality of syllabus. Basically we have considered the posts or tweets of engineering students regarding the problems they face in their academic experiences majorly because: Engineering colleges are facing problem of retention. Engineering studies represent a major part in future manpower and have an impact on nation’s economy. Supported understanding of problems and issues in students’ life because of the huge engineering syllabus, educators will create additional hip to choices on correct interventions and services that will open the door for students and help overcome barriers in learning. Twitter may be a standard social media website. Its content is generally common and extremely epigrammatic (no quite one hundred forty characters per tweet). Therefore, we are analyzing post on twitter by students.

II. LITERATURE REVIEW

Study overview the analysis of data on social media using algorithms and this article will help us for qualitative analysis and data mining analysis and Methodology used Multi label classification algorithms. Mainly there are two data-driven approaches emerging in education viz. Learning analytics and educational data mining (EDM). These will help us to know the students problem that they face in education by analyzing the data and then we can give these records to administrator to take appropriate decision according to the class in which students face more problems. Another technique i.e. EDM helps to know the problems regarding research.
III. PROPOSED SYSTEM

a) Mining Twitter Data

To produce the knowledge of different subject, Researchers selected the twitter content for analysis. Qualitative content analysis and linguistic analysis are two methods which used in these studies. Usually on statistical model and algorithms these studies have a greater significance. Some majorly used algorithms are SVM i.e. support vector machine, M3L i.e. Max-Margin multi-label classification and naïve bayes classification. As per the number of classes present in the algorithm, there are mainly two types of classification approaches are there, viz. binary classification and multi-class classification approaches. Only two classes are there in binary classification, while multi-class classification involves more than two classes. Both classification are single-label classification systems. Single level classification means each tweet will be considered only in one class. In Multi-label classification, allows each tweet to fall into several classes at the same time. Sentiment analysis won’t work here as it doesn’t give much knowledge about students actual problem so as to provide relevant interventions and services for students. Basically our main aim is to know the students problem well and classify it properly. Even it is not possible sometimes to determine the tweet that in which class it will fall. Sentiment analysis is different then our study. In our study, we use classification technique where we allow each tweet to fall into multiple categories at the same time. As it contains the keywords of more than two classes then it will categorize into more than one class.

b) Work flow of Project

c) Tweet Collection

As students uses different languages while posting their tweets it is beat difficult to analyse and also some of them use special characters. There will be also many tweets where students might tweet sarcastically which means the tweet written will have some meaning but when understood properly it will have a different meaning. Now in order to collect the tweets, the Twitter APIs[1] can be used to obtain the second data set. We will search with the keywords like engineer, students, problems, class, homework, late night, lab, etc. There also some other words like such as #ladyEngineer, #engineeringMajors, #hugeenggsyllabus, #engineeringstudent, #collegeProblems, #sleeplessnights, #assignments, and #homeworkproblem and many more. This words with hashtags can be used to search in database. Since there were no predefined categories for the problems faced, we needed to identify what the students were trying to say in every tweet. Thus, we will start with engineeringproblems data set In order to manually analyze the text content Inductive content analysis is a popular research method.

1. Development of Categories

Nowadays many engineering students face different problems and it becomes difficult for them to study for the exams properly. Now few of the common problem that the students face include heavy study load, sleep problems, depression, frustration, gender inequality, lack of social engagement, staying shy and don’t talk much about topics to other students, feeling stress, motivation is not given properly, problems i.e. health problem, and others. These is used classify tweets without any issues. Ahead are the five major categories which were found as a problem and root cause for other problems: heavy study load like students having problems with too much syllabus, lack of social engagement: where students fails to contribute in social activities, negative emotion when students are not feeling well or disturb with the education process, sleep problems this is the general one that many students might be facing, and diversity issues. So problems of students will be categorize into these categories. It’s not that engineering is only bad, there will be many good things about it too and there will be positive posts too with #engineeringperks and this will be classified into ‘Good Things’ category[5]. We can also have a thing while doing this project that there will be many tweets that might fall into more than one category. Thus we require multi-label classification approach as it is different from single label in which tweet can be classified into any particular category. Each and every category has many keywords and it is possible that one keyword might fall into multiple categories.

d) Data Preprocessing

Data preprocessing is the most important step in this project. The data with which we will be dealing will be raw data and it will be like not properly formed data. There will be multiple alphabets in a particular wordeg. “hellllooooo”,

![Fig 1. Workflow](image-url)
“hateestudyyyyy” etc. This data preprocessing will help us to identify the data and also make the data readable. In this we will just be making use of simple data mining steps.

1. Data Reduction

Now, whatever tweets we get they are in large numbers and it becomes very difficult to read those tweets and segregate them. There will also be many irrelevant tweets meaning that these tweets will not be related to our project, so we will be eliminating those tweets. Ex. “I just want to run away from my problems”. This tweet tells us that the person is facing problem in his life and it is not related to engineering and thus we eliminate this tweet.

2. Data Transformation

Whatever tweets that we collect are basically made by the students and obviously they don’t have this habit of making simple formal tweet. Thus we can say that these tweets are not well formed, so we will be smoothing them.

e) Naïve Bayes Multi-Label Classifier

Nowadays there are many popular data mining techniques that are used but we will be considering Naïve Baye’s Multi-label classifier. We found Naïve Bayes classifier to be very effective on our data set [4]. One of the best way to implement this is to transfer it to multiple single label classifier [2]. Tweets can be classified into more than one category. Multi-label Naïve Baye’s classifier’s basic procedure is given below. Suppose there are N no. of words in the data \( W = \{w, w_2, \ldots, w_N\} \), and a total number of L categories \( C = \{c_1, c_2, \ldots, c_L\} \). If a word \( w_n \) appears in a category \( c \) for \( m_{w_n c} \) times, and appear in categories other than \( c \) for \( m_{w_n \text{c}\neq c} \) times, then probability is calculated as given below:

\[
p(w_n | c) = \frac{m_{w_n c}}{\sum_{i=1}^{n} m_{w_i c}}
\]

Learning Methods

In this project we will also consider the tweets of how he/she studies i.e by lectures or by e-learning. Nowadays, it’s not necessary that all the students are not able to concentrate during the lectures and they find it boring just sitting at one place, and then on the other hand there are students who find E-learning pretty good as compared to lectures as it involves great interaction and has slide that will help the students grasp the topic fast as compared to lectures.

f) Comparing and generating result

Along with the classification of the tweets in the problems faced by the students, learning techniques will also be classified. Data collection on the basis of this classification will be done and further it will be classified. The result of the classification of the students’ problem area and the result of the learning techniques will be divided in the ratio of E-learning and classroom lectures. For eg. If we get 800 tweets for heavy study load category and ratio of issues in E-learning and classroom lectures is 1:3, then this ratio tells us that the tweets are divided into 200 for issues in E-learning where students feel heavy study load and 600 for the issues in Lectures.

IV. OBEE Model

Fig 2. OBEE Model

a) Build Layer

This layer is the base of entire architecture, and hence the name. Build layer forms course objectives for the students from the lesson objectives. Basically every teacher will teach every concept mentioned in the course objectives. Based on the tests taken in the assessment layer the students will be classified into Excellent, average and Poor category.
b) Assessment Layer

This is the second layer of 5 layer OBEE architecture which deals with the two aspects: Content Delivery and Assessment Tools.

- Content Delivery: There are various ways to deliver contents. Some of them are Class Room Teaching, Tutorials, Lab Experiment, Self-Learning Online Resources, Slides/Handouts/Notes and Seminars.

- Assessment Tools: This aspect is the heart of Assessment layer. After each delivery of course by the course teacher, assessment of the student has to be taken to categorize them based on their understanding level.

c) Feedback Layer

It is widely recognized that feedback is an important part of the learning cycle. This layer is all about the feedback (course exit survey). The central question is why feedback is necessary? The answer has 4 key points:

- Feedback is very essential to formative assessment.
- Feedback creates learner autonomy.
- Feedback goes parallel with teaching-learning process.
- Feedback directly affects student as well as teacher experience.

d) Target Layer

Mapping the course outcomes with the program outcomes is what the Target layer is all about. It may be possible that one particular course outcome may match with two or more program outcomes. It is needless to say that, target layer assessment cannot come under assessment layer as it does not assess the individual course outcomes for each student.

e) Management Layer

Time management layer is integrated layer in 5 layer OBEE model. It ensures that all the academic activities are happening synchronously and make sure that preparation and delivery of outcomes, assessment, feedback, mapping is done on the stipulated period of time.

---

**Fig 3. Merging categories with OBEE**

We will be referring the OBEE (Outcome Based Engineering Education) Model. Basically it will consist of 5 layers and each of them consisting of the given content. We will be mainly focusing on the feedback layer as this helps us know what the students actually want and what change they want in the system. The target line is the phase between Feedback Layer and Target Layer. Based on the POOR understanding level of students, remedial classroom teaching can be initiated to achieve common target line. Feedback layer would be presented in form of spreadsheets to know what students actually feel in order to avoid paper work. The architecture consist of 5 layers namely Build layer, Assessment layer, Feedback layer, Target layer and Management layer. Build layer provides the direction towards development of Course Objectives. Through content delivery, teacher can assess the students using various assessment tools. A feedback can be taken from the students in order to enhance the quality of the course. Here, feedback in our project is in the form of tweets. In the feedback layer of the OBEE Model, the feedback of the students that is nothing but the problems that are faced by them during their learning will be merged and it will be sent to the management layer so that they can take appropriate decisions in improving the education quality and making it better for the students and lessen their problems even by a small amount which will be a very great thing.

1. Limitations and Conclusion
Using different methods like large scale data mining technique and qualitative methods, this study explores uninstrumented space on social media application to study and understand engineering students’ experience. Friendly relation between analysis done in the tweets and algorithms for classification will be provided by this [3]. This work will basically help us to take proper actions and get insights into students’ problems and help us to take proper action. But, there are some disadvantages of this study are: Firstly, it is not possible that all the students share their views about education on social media. Also, this being a category of problem faced we might get few tweets and also many re-tweets. Secondly, even if there will be certain tweets having keywords in it like ‘study’ or ‘problems’, many irrelevant tweets where some person will tweet about his personal life problems or a quote to get over problems, such tweets will also be considered and sometimes it won’t count all tweets that are related. Thirdly, in this paper we collect tweets only about engineering and not considering other domain like Medical, Law, Commerce etc. There will be students in these domains too who might be facing these problems. Other possible future work could analyze students’ generated content other than texts (e.g., images and videos), on social media sites other than Twitter (e.g. Facebook, Tumbler, and YouTube). Future work can be done on why and how student seek and used social media site. This study will help administrators to understand the engineering students’ college problems and take proper actions to improve the education quality. Our study will provide benefits to education administrator in learning analytics, mining the educational data, and learning technologies. Major limitations will be removed and education level will be increased by our study and it will help students. It is cleared that after implementing proposed work many of the useful tweet which were going under the others and not considered. This types of tweets we will be collecting and classify them into another category like “Good Thing”. Finally we will get the response from the students to solve their educational problems and result will be generated by the graph.

References