Analyzing and Mining Social Media Data for Predicting Student's Learning Experiences

Chandu Asok
PG Scholar
Computer Science and Engineering
Younus College of Engineering and Technology,
Kollam, India-691010

Rageena M
Assistant Professor
Computer Science and Engineering
Younus College of Engineering and Technology,
Kollam, India-691010

***Abstract*—Recently, social media is playing a vital role in social networking and sharing of data. Social media is favored by many users as it is available to millions of people without any limitations to share their opinions, educational learning experience and concerns via their status. Students posts on social network gives us a better concern to take decision about the particular education systems learning process of the system. Evaluating such data in social network is quite a challenging process. In the proposed system, there will be a work ow to mine the data which integrates both qualitative analysis and large scale data mining technique. Based on the different prominent themes tweets will be categorized into different groups.**

Nave Bayes classifier will be implemented on mined data for qualitative analysis purpose to get the deeper understanding of the data. It uses multi label classification technique as each label falls into different categories and all the attributes are independent to each other.

Index Terms—Social Network, Twitter, Education, Nave Bayes, Multi-label Classification, Data mining.

1. INTRODUCTION

Social media has become one of the comfortable medium for people to share their feelings instinctively. According to the survey sharing of the data is high in the social sites like twitter and facebook. Students pay more attention to share their feelings spontaneously in a relaxed, informal environment more than the formal classroom environment [4]. An information educational researcher can easily understand students experience from outside of the class and they can get full details about the education system.It is very helpful for an institution to understand the difficulties of the student he/she facing in the learning system. However this data or information is used to upgrade the education system of the particular institution. These social network data mining provide an opportunity to make changes in education system ultimately to make an impact on economic growth as students play a vital role in the future workforce.

The research goal of this learning are:-

1) To show a method of social media information mining for educational reasons, combining the techniques of both qualitative analysis and with the suitable data mining techniques.

2) To understand students informal conversations on Twitter, in order to predict the problem faced by them.

Here the focus is mainly on the student's tweets in which they post on the twitter from which the problems faced by the student is being identified most suitably in an informal way of communication. This study differentiates the formal and informal way of communication. The informal way of communication can be the method of conducting surveys, questionnaire etc.

This study can be very useful because:

1. Most of the schools as well as the colleges have being struggling to find out the problems faced by them, since the students constitute as a part of nations's workforce and have a significant impact on nation's economic growth.

2. After getting out the problems the educators can make significant services to overcome the barriers faced.

Educational researches have been using old ways such as surveys, interviews, focus groups, classroom activities to collect data related to students learning experiences. That will reducing scalability problem. As optimistic about their experiences, students need to reflect on what they were thinking and doing sometime in the past, which may have become obscured over time. There is no research found to directly mine and analyze student-posted content with considering the students problem from uninhibited spaces on the social web with the clear goal of understanding students learning experiences. In this paper several tweets are collected based on # engineeringProblem, #nerdstatus,

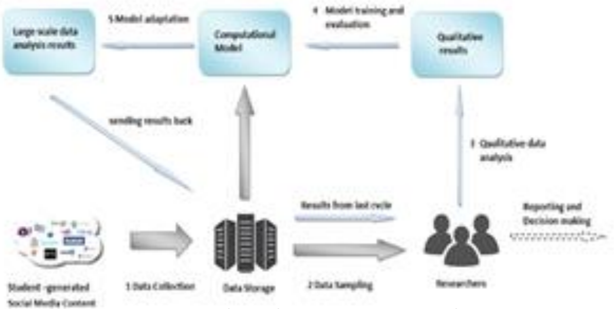

Large scale data analysis results
5 Model adaptation
Computational Model
4 Model training and evaluation
Qualitative results
sending results back
3 Qualitative data analysis
Results from last cycle
Reporting and Decision making
Student -generated Social Media Content
1 Data Collection
Data Storage
2 Data Sampling
Researchers

Figure 1. Overall work flow for making sense of social media data integrates qualitative analysis and data mining algorithms.

#studyproblems and tweets based on top engineering colleges in india. These helps in describing the process to locate the relevant data and relevant Twitter hashtags (a Twitter hashtag is a word beginning with a # sign, used to emphasize or tag a topic).

The workflow which includes both qualitative analysis and data mining algorithms are developed in order to improve the performance and can be considered to be an iterative cycle. In the figure, the width of gray arrows represents data volumes wider indicates more data volume. Light gray arrows represent data analysis, computation, and results flow. The dashed arrows represent the parts that do not concern the central work of this paper.

The major steps are:

Step 1: Data is being collected from the student generated social media contents.

Step 2: Researchers conducted an inductive content analysis on samples of the # engineeringProblems dataset. Step 3: A detailed data analysis is done.

Step 4: Major problems of engineering students are encountered are categorized into several categories and a multi-label classifier is implemented by Naive Bayes. algorithm.

Step 5: The classification algorithm is used to train a detector that could assist detection.

The rest of the paper is organized as follows: Section 2 describes the literature survey to the paper. Section 3 describes the implementation method in detail, section 4 gives the result, section 5 gives the conclusion and finally section 6 gives the future work.

2. LITERATURE SURVEY

With the intensification of the Internet communiqu techniques, the World Wide Web have became a extremely imperative display place for users for a interaction with each other. In the course of this display place, users could share with a ease and broaden in sequence and dreams to any one and from anywhere. Twitter is an exceptionally trendy micro blogging site, where usersrummage around for appropriate and social information referred as breaking news, posts about celebrities, and trending topics. Users send short transcript messages referred as tweets, have limitation of 140 characters by length and could be view by users followers. A follower is any person called when he is choosing to have others tweets posted on ones timeline. For real-time, Twitters have been worn as a channel information giving out and it has been used in various brand elections, campaigns, and likely to a news media.

The data mining done on social media data covers many uncover features of the social media or the social web i.e. Twitter, Facebook and Youtube. Mining file containing text has been known as Intelligent Text breakdown or acquaintance discovery in Text or it can said that Text Data Mining which can be used to mine the social media data. Mostly the social media data are unstructured format and to retrieve information from that is complex due the massive information. So, it requires specific processing methods and algorithms to extract useful information from that social web data.

One of the major research projects regarding engineering students experiences is the Academic Pathways Study (APS) conducted by the Center for the Advancement of Engineering Education (CAEE) [1]. APS consists a series of longitudinal and multi-institutional studies on undergraduate engineering students learning experiences and their evolution to work. They used various research methods including surveys,structured interviews; semi-structured interviews, engineering design task, and small focus groups. The CAEE website presents research briefs from the APS study including topics such as developing identity as an engineer, conceptions of engineering, workload and life balance, and persistence in engineering as a college major and as a career.

Educational Data Mining is an promising regulation, concerned with budding techniques for discovers the exclusive types of data that come from educational background, and using those techniques to better understand students, and the settings which they studied in.. Learning analytics and educational data mining (EDM) [2] are data-driven approaches emerging in education. These approaches analyze data generated in educational settings to understand students and their learning environments in order to inform institutional decision-making. Educational Data Mining (EDM) is the application of Data Mining (DM) techniques to, its objective is to examine these type of information in order to determine educational research problems.

The Naive Bayes Classifier technique has been based using Bayesian theorem and the best wellmatched to the dimensionality of the inputs has sky-scraping. Over and over again outperforms more sophisticated classification methods, yet it is simple to operate. For the models, Maximum Likelihood estimates all the parameters. To estimate the parameters, it has been require of small number of training. It operates glowing and powerfully in supervised learning.

3. IMPLEMENTATION

This study built a multilabel classifier to categorize tweets stands on the categories developed in content analysis phase. There are numerous well-liked classifiers generally used in data mining and machine learning field. It establishes that Nave Bayes classifier to be very efficient for this dataset compared with further multilabel classifiers.

3.1. Data Collection

It is challenging to collect social media data related to students experiences because of the irregularity and diversity of the language used. The searching of data is conducted using an educational account on a commercial social media monitoring tool named Radian6. The Twitter APIs can also be configured to accomplish this task. The search process was exploratory. The searching was conducted based on different boolean combinations of possible keywords such as engineer, students,campus, class,homework, professor, and lab, since the main aim is to solve the problems faced by the Engineering Student's.

A multi label classification were arises in which many tweets belongs to more than one category which conflicts the concept of single label classification in which one tweet belong to one category. The label or label set is the category to which one tweet belongs to. Some of the measures such as Cohens Kappa, Scotts Pi, Fleiss Kappa, and Krippendorfs Alpha [6], [7] is used for data that belong to single label classification and cannot be used in multi label classification. Therefore here it uses F1 measure which is the harmonic mean between two datasets. When the two sets of data are exactly the same, then the F 1 score will be 1 and if the two sets of data are completely different then the score will be 0. Thus it represents how close two label sets are assigned.

$$F_i = \frac{1}{N}\sum_{i=1}^{N}\frac{2p_1 i.p_2 i}{p_1 i + p_2 i}$$

3.2. Text Pre-processing

Many symbols are being used by Twitter to convey special meaning. For example, # is used to indicate a hashtag, @ is used to indicate a user account, and RT is used to indicate a re-tweet. Stop words a, an, and, of, he, she, it, non letter symbols, and punctuation also bring noise to the text. Thus the text are preprocessed before training the classifier.

1. The #engineeringProblems hashtags were removed and for other hashtags only # sign is removed and the hashtag texts were kept as such.

2. In order to detect negative emotion and issues negative words are used. Thus negtoken was substituted for all the words ending with nt and other common negative words (e.g. nothing, never, none, cannot).

3. Non-letter symbols and punctuation contained words are removed which includes the removal of @ and http links and RTs.

4. For repeating letters in words, If any detection of more than two identical letters repeating, then a replacing will take place with one letter. Therefore, huuungryyy and sooo were corrected to hungry and so. muuchh was kept as muuchh. Originally correct words such as too and sleep were kept as they were.

5. The common stop words were removed by using information retrieval toolkit. We kept words like much, more, all, always, still, only, because the tweets frequently use these words to express extent.

3.3. Naive Bayes Multi-label Classifier

The Naive Bayes classifier [5] is a straightforward probabilistic classifier which is based on Bayes theorem with strong and nave self-government assumptions. It is one of the most basic text categorization method with various applications in email spam exposure, privatemail sorting, document categorization, , language discovery and sentiment discovery. Naive Bayes executes well in many difficult real-world troubles. Even though it is frequently outperformed by other techniques such as boosted trees, Max Entropy, Support Vector Machines etc, Naive Bayes classifier is extremely efficient since it is less computationally and it requires a small amount of preparation information.

This classifier considers each sub words in the review and accordingly classifies the reviews in different categories. Let S is the Sentence

Step 1: Define categories c=c1,c2,c3,...,cn

Step 2: Read data from a database.

Step 3: Divide S into sub worksw1,w2,w3wn split.

Step 4: Check sub words w1,w2,w3..wn for every categories.

Step 5: If words match with categories c1,c2.c3.cn increment the counter for that categories.

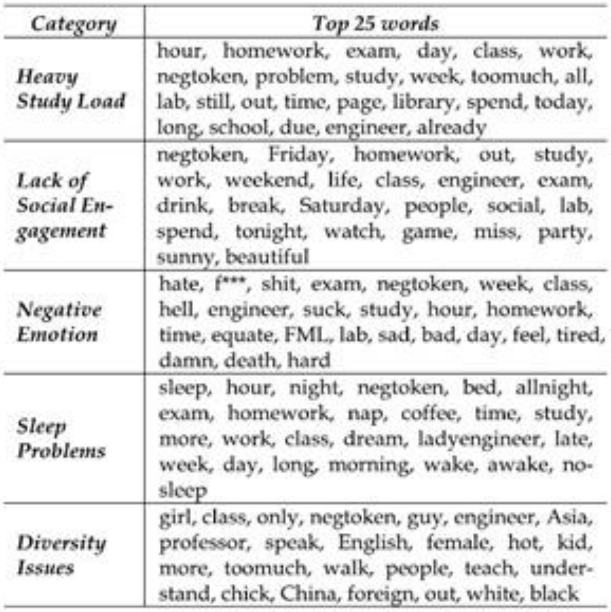
Step 6: Find probability of each category.

3.4. Inductive Content Analysis

Because social media content like tweets contain a large amount of informal language, sarcasm, acronyms, and misspellings, meaning is often ambiguous and subject to human interpretation. Rost et. al argue that in large scale social media data analysis, faulty assumptions are likely to arise if automatic algorithms are used without taking a qualitative look at the data. For example, LDA (Latent Dirichlet Allocation) is a popular topic modeling algorithm that can detect general topics from very large scale data. LDA has only produced meaningless word groups from our data with a lot of overlapping words across different topics.

There were no pre-defined categories of the data, so we needed to explore what students were saying in the tweets. Thus first of all an inductive content analysis was performed on the #engineeringProblems dataset. Inductive content analysis is one popular qualitative research method for manually analyzing text content. Three researchers collaborated on the content analysis process.

3.5. Categories



Category	*Top 25 words*
Heavy Study Load	hour, homework, exam, day, class, work, negtoken, problem, study, week, toomuch, all, lab, still, out, time, page, library, spend, today, long, school, due, engineer, already
Lack of Social Engagement	negtoken, Friday, homework, out, study, work, weekend, life, class, engineer, exam, drink, break, Saturday, people, social, lab, spend, tonight, watch, game, miss, party, sunny, beautiful
Negative Emotion	hate, f***, shit, exam, negtoken, week, class, hell, engineer, suck, study, hour, homework, time, equate, FML, lab, sad, bad, day, feel, tired, damn, death, hard
Sleep Problems	sleep, hour, night, negtoken, bed, allnight, exam, homework, nap, coffee, time, study, more, work, class, dream, ladyengineer, late, week, day, long, morning, wake, awake, nosleep
Diversity Issues	girl, class, only, negtoken, guy, engineer, Asia, professor, speak, English, female, hot, kid, more, toomuch, walk, people, teach, understand, chick, China, foreign, out, white, black

Figure 2. Categories

The Researchers have found out five different prominent categories or themes from the research they have conducted. The five dfferent Categories are:

Heavy Study load :Students are not able to handle the stressful life as it leads to lack of social engagement, lack of sleep, stress, depression, and some health problems. Previous studies also show that engineering students require more balanced life than their academic environment [8]. For example: going to bed at 3 A.M. Still have about 8 hrs of homework and studying to do. . .

Lack of Social Engagement :Social engagement in students helps the students in releasing the stress and therefore the students must be involved in doing the social works. Lack of the social works can result various problems among students which in turn result in anti social image of engineers. The students must sacrifice the time for enjoying the holidays special occasions with family and friends etc. Thus for building up an efficient engineer the students must engage in doing social engagements which is a beneficial for learning Some students embrace the anti-social image, while most others desire more social life [9]. In short we can say that the society needs students or engineers who has the capability to handle different situations and work with the people to solve the problems [10], [11]. For example: Chemistry and calculus homework everyday of thanksgiving break.

Negative Emotion :A negative emotion is being categorised only when emotions such as hatred, anger, stress, sickness, depression, disappointment, and despair were identified. There are a lot of negative emotions flowing in the tweets. The hashtag #engineeringProblems helps in determining these emotions. These emotions can arise for example when a student gets stressed with his/her homework/schoolwork. Therefore it is very much important to manage the students psychological emotions and stress.
For example: is it bad that before I started studying for my tests today that I considered throwing myself in front of a moving car??.

Sleep Problems :Lack of sleep which is widely seen in engineering students, results in many psychological, psychological and physical health problems. This arises because of heavy study load and stress. This is one of the important issues that are to be considered.
For example: I wake up from a nightmare where I didnt finish my physics lab on time.

Diversity Issues :One of the diversity issues is the lack of female students for engineering and this can result in the bad character towards female students by the male students in engineering as they do not get opportunity to mingle with the female engineering students. Another diversity issues include the problem of understanding the lectures of foreign professors in the class. The students find difficult in adjusting with their culture and many behaviours.
For example: Lets start with an example, tell me something you know nothing about Professor . . . girls. Students. lol.

3.6. Updation

If some grammatical forms of a particular word present in a category are active in a particular tweet taken. Then the classifier does not recognize that this corresponding word belongs to that category.
For example the category heavy study load has the presence of a particular keyword study. If in the inut tweet suppose the word retrieved is studies or studying or studied, then the classifier will not recognize that it belongs to the category heavy study load. This process can limit the predicting quality.
So inorder to avoid such a problem to takes place we are updating the grammatical forms of a particular word with the suitable word present im the category. In the above example if we find any word such as studies, studying, studied; then these words will replaced by the word study, since study is present in the category heavy study load. Through this process the prediction accuracy can be improved and has the capability of getting good result when compared to other.

4. RESULTS

In this study, through a qualitative content analysis, found that engineering students are largely struggling with the heavy study load, and are not able to manage it successfully. Heavy study load leads to many consequences including lack of social engagement, sleep problems and other psychological and physical health problems.

Building on top of the qualitative insights, we implemented and evaluated a multi-label classifier to detect engineering student problems. This work is only the first step towards revealing actionable insights from student generated content on social media in order to improve education quality. The result is shown in the figure.

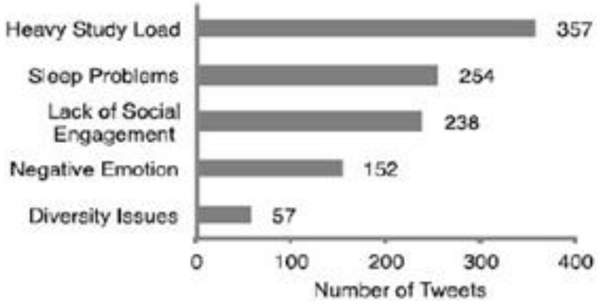


Figure 3. Result

5. CONCLUSION

These studies overcome the limitations of a manual qualitative analysis and large scale computational analysis of user generated textual content. It helps the researchers in learning analytics, educational data mining, and learning technologies. This analyzes social media data for educational purposes which can throw light to educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students learning related experiences.

6. FUTURE WORK

Apart from the work done towards this system, future work mainly comprises of the following objectives:

- In future we can collect large user generated data other than texts which may include videos and images for analyzing the user experience with exact results. Graphs are used to show the both positive and negative experience results on a yearly basis.
- Building a tool based on the social media data and the user social sites performance.

REFERENCES

[1] M. Clark and K. Smith, Academic pathways study: Processes and realities, in Proceedings of the American Society for Engineering Education Annual Conference and Exposition, 2008.

[2] 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. 317, 2009.

[3] 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. 317, 2009.

[4] Pallavi K. Pagare, Analyzing Social Media Data for Understanding Students Problem, International Journal of Computer Applications (0975 8887).

[5] Payal S.Jain and Pallavi S.Panhale, EFFECTIVE MINING SOCIAL MEDIA DATA FOR UNDERSTANDING STUDENTS LEARNING EX- PERIENCES, International Research Journal of Engineering and Technology (IRJET)Volume: 03 Issue: 01,Jan-2016.

[6] K. Krippendorff, Reliability in content analysis, Human Communica- tion Research, vol. 30, no. 3, pp. 411433, 2004.

[7] M. Lombard,J. Snyder-Duch, and C. C. Bracken, Content analysis in mass communication: Assessment and reporting of intercoder relia- bility, Human communication research, vol. 28, no. 4, pp. 587604, 2006.

[8] H. Loshbaugh, T. Hoeglund, R. Streveler, and and K. Breaux, Engi- neering School, Life Balance, and the Student Experience, presented at the ASEE Annual Conference and Exposition, Chicago, Illinois, 2006.

[9] H. Loshbaugh,and B. Claar, Geeks are chic: Cultural identity and engineering students pathways to the profession, in Proc. ASEE, 2007.

[10] ABET, ABET - Criteria for accrediting engineering programs, 2012- 2013 2013-2012. [Online], Available: http://www.abet.org/engineering-criteria-2012-2013/. [Accessed: 10-Sep-2015]. National Academy of Engineering The engineer of 2020: Visions of engineering in the new century. Washington, D.C.: National Academies Press, 2004