A Survey on Students’ Analysis through Social Media
Esha Kadam¹, Shraddha Joshi ², Sayali Joshi ³, Amruta Jape⁴, Prof. Pardeshi N.G⁵
Final year students¹,²,³,⁴,⁵
Computer Engineering Department
Sanjivani College of Engineering, Kopargaon, India

Abstract:
An increased use of social media sites has been observed by students pursuing higher education. All this sites like twitter, Facebook, google+ etc. helps user to keep in touch with friends, relatives, neighbors, colleagues etc. Engineering students are found making profound use of this social media sites to express their views without peer pressure or staff pressure. The data available on these sites may be present in unstructured or structured format and this data provides valuable information related to students learning experience and is useful to improve or enhance the education system followed by engineering institutes. This paper describes how tweeter data can be used by education analysts to know the learning experience of engineering students. We build a multi-label classifier to classify tweets into some pre-defined classes. Several popular classifiers are widely used in data mining domain. We found that Naive Bayes classifier very effective as compared with other multi-label classifiers. By using Naive Bayes classifier, we are able to identify student’s learning experience properly.

Index Words: Naive Bayes classifier, Stop-words, Social networking, Twitter, Term frequency

I. INTRODUCTION
Aim of our project is to understand the learning experience of engineering students through the tweets they post on twitter. We are going to use few techniques of data mining in order to achieve the goal decided. Based on the tweets made by the students it will be possible to education analyst to see whether what problems students are facing due to present schema of teaching and learning. The problems majorly focused are ‘heavy study load, lack of social engagement, sleeping problem, diversity issues, and negative emotions’. We are going to explore engineering student’s informal tweets without causing any privacy and security issues about their credentials.

Twitter is chosen as the platform for analyzing because of following reasons:
1. It is popular social media site.
2. Its content is mostly public and concise (only 140 characters per tweet are allowed).
3. Free APIs are provided by twitter in order to stream data.

Naïve Bayes classification algorithm is to be used in these concepts as we found that Naïve Bayes classifier to be very effective on dataset compared with other classifiers. Social media sites such as Twitter, Facebook and YouTube provide great platform for students to share joy and struggle, various emotions and stress. In daily routine due to reasons like peer pressure or teachers pressure students hesitate to express their opinion about what are their views about their day to day experience. On other hand on various social media sites, students discuss and share their everyday encounters in an informal and casual manner. Students’ digital footprints provide a whole new perspective for education researchers and practitioners to understand students’ experiences outside the controlled classroom environment. This understanding can inform the policy makers on interventions for at-risk students, improvement of education quality. This all will thus enhance student recruitment, retention and success.

Large numbers of students are pursuing engineering in Pune University. The locations of engineering students are varying. If one decides to perform pure manual analysis to understand engineering students learning experience it a next to impossible task due to above mentioned reason. Again pure manual analysis cannot deal with over growing scale of data. Education researchers have been using methods such as surveys, interviews, focus groups, classroom activities to gather data related to students learning experience. These all methodologies are time consuming and can’t be repeated when the volume of data increases. Additionally students may hesitate to express their feelings in front of any other human. To best of our knowledge, there is no technique found to directly mine and analyze student posted contents for understanding students learning experience. The workflow developed for understanding social media data integrates qualitative analysis and data mining algorithms.

The research goals of this study are:
1. To demonstrate how social media data can help educational analysts to understand problems of engineering students. We are integrating both qualitative analysis and large-scale data mining techniques.
2. To delve into engineering students’ informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences. We focused on engineering students’ posts on Twitter about problems in their educational experiences mainly because:
   1. Engineering schools and departments are struggling with student recruitment and retention issues.
2. Engineering students constitute a significant part of the nation’s future workforce and have a direct impact on the nation’s economic growth and global competency.

3. Based on understanding of problems and issues in students’ life, policymakers and educators can make more proper decisions that can help students overcome barriers in students learning experience.

II. LITERATURE REVIEW

In order to propose our system we have gone through already existing systems. A brief review of all other papers we referred is mentioned below:

1. “Celebrity’s self-disclosure on Twitter and parasocial relationships: A mediating role of social presence”- Jihyun Kim, Hayeon Song:
   This paper focuses on how self-disclosure on twitter affects fans views about celebrity. They conducted an online survey among a sample of 429 celebrity followers on Twitter. Result obtained were that celebrities’ professional self-disclosure, personal self-disclosure, and fans retweeting behavior increased fans feeling of social presence, hence positively affecting parasocial interactions with celebrities.

2. “Utilizing social media data for pharmacovigilance: A review”- AbindeSarker, Rachel Ginn, Azadeh Nikfarjam, Karen O’Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadahya, Graciela Gonzalez:
   The main objective of this paper was Automatic Monitoring of Adverse Drug Reactions (ADR). In this paper, they performed a methodical review to characterize the different approaches to ADR detection/extraction from social media, and their applicability to pharmacovigilance. Additionally, they presented a potential systematic pathway to ADR monitoring from social media.

3. “Naive multi-label classification of YouTube comments using comparative opinion mining”- Asadullah Rafiq Khan, Madiha Khan, Mohammad Badruddin Khan:
   In this paper, the authors have used Naive Bayes machine learning algorithm to perform multi label classification. They find out the sentiments of the commenters for different options.

4. “Social learning analytics”- Simon Buckingham Shum and Rebecca Ferguson:
   In this paper, they proposed that the design and implementation of effective Social Learning Analytics present significant challenges and opportunities for researchers and enterprise. They concluded by revisiting the drivers and trends, and considered future scenarios that they may see unfold as SLA tools and services mature.

5. “Interpreting Students behavior using opinion mining”- Pooja R Takle, Prof. Narendra Gawai:
   This paper describes analysis and classification of structured data by using various classification algorithms. According to their opinion Naive Bayes Multi-classifier is used. Naive Bayes gives good results but it’s very consuming. So to overcome this limitation they proposed a technique called as “Memetic Classifier” based on genetic algorithm. They did a comparative study of classification techniques such as Iterative dichotomiser, Naive Bayes and memetic classifier using common dataset.

6. Using social data for understanding students experience and decision crisis correlation”- Thahin Taj A’, Dr. R. Bahu:
   In this paper, they developed a workflow to combine both qualitative analysis and large-scale data mining techniques. They focused on students’ Twitter posts to understand issues they face in their daily life. They found that engineering students face problems such as heavy study load, lack of social engagement, sleep problems, diversity issues and negative emotions mainly. Based on these results, they implemented a classification algorithm to classify tweets reflecting problems faced by students.

III PROPOSED SYSTEM

The goal of the proposed system is to find out what problems are faced by the engineering students, by gathering and analyzing their conversations made on Twitter. Further upon understanding their problems we aim that the educational analysts, policy-makers would make the changes in the curriculum of the engineering so that students could help themselves in overcoming the hurdles faced by them in their learning. Our system will classify the problems faced by students in 5 different classes viz, heavy study load, negative emotion, lack of social engagement, diversity issues, sleep problem.

The system so designed will have the following modules:

- Tweet Collection.
- Preprocessing.
- Training.
- Prediction.

- Tweet Collection:
   Students post their views about everything that they experience on Twitter. May it be their exams, study load, homework, practical, lectures and many more. To collect the tweets from twitter several tools such as Radian6 are available. Besides Twitter also provide Twitter API for the same. We are using the Twitter API in our systems to retrieve the tweets needed for the analysis of the students.
• Pre-processing:
However, the tweets that are so fetched cannot be directly used for further work. This is because it may contain many words which may be of no use or which may need some kind of processing. For example in some words, letters are repeated like “boringgggg”, “helloo00”, “c00ool” etc. This is done to express the strong feelings. Also, #sign which precedes the tags on twitter is not required. This is also applied to @ and other punctuation marks. Pre-processing involves the following steps:

1. Tokenization
2. Filtering
3. Applying TF

1. Tokenization:
It means to form tokens of a tweet i.e. to separate every single word of the tweet.

2. Filtering:
It is used to solve the problems mentioned above and also to remove the stop words. Example of stop words are: ‘a’, ‘is’, which don’t have any meaning.
If required, we are also going to use stemmer library to do stemming of the words. Stemming means if there are words say ‘going’ ,’goes’, ’gone’ they will be stemmed to ‘go’.

3. Term Frequency:
It will help to determine how frequently a word is used in tweets.

• Training:
Naive Bayes multi-label classifier is used for training phase. A multi-label classifier is used because a single tweet may fall in more than one classes. So in case if a tweet fall in more than one class, it is included in all those classes. Hence a multi-label classifier is used. Also we are providing an admin login to our system who can add keywords to the five mentioned classes at runtime.

• Prediction:
When a user as in a educational analyst logs into the system, dynamically tweets would be downloaded using the consumer key of ours. Tweets so downloaded will be stored in a file using the concept of serialization. Preprocessing will then be applied over these tweets. Training set is used in the prediction phase to find the probability of each of the five classes with respect to individual tweet collected dynamically. Bayes theorem is used for the same. Depending upon which class i.e heavy study load, lack of social engagement, diversity issues, sleep problem, negative emotion , has the highest probability, the corresponding problem will be treated to be faced by the majority of the engineering students.

Output is displayed to user in two forms i.e in graphical format and as numerical count. Thus the result of the system will help educational analysts to refine the curriculum of students.

CONCLUSION
Our study is helpful to researchers in learning analytics and education policy makers. It provides a workflow for analyzing social media content for educational purposes that overcomes the major fallouts of manual analysis. Our study can inform educational administrators, practitioners and other relevant policy makers to gain further understanding of engineering student’s learning experiences. This paper describes how mining is better over traditional manual analysis. This system is implemented with a Naive Bayes classifier which outperforms other multi-label classifiers.

As per the survey done this system is very useful in institutes, organization and universities. This system is also very useful in institutes, organization, banking domain, government sectors etc. in future for identification of employee’s actions their behaviors, product feedback and other services.

As per the survey and work done this system is very useful in institutes,organizations and universities. This system is also very useful for industry, manufacturing companies, banking sectors, government sectors etc. in future for identification of employees actions, their behaviors, product feedback, for banking feedback and related to their services.

REFERENCES
4. Simon Buckingham Shum and Rebecca Ferguson.”Social learning analytics, educational technology and society,15(3),3-26
5. Pooja R Takle, Prof. Narendra Gawai, ”Interpreting Students behaviour using opinion mining” International Journal of innovative research in computer and communication engineering.