Naive Bayes Classifier for web text Analysis

Kasat Priyanka[1], Dr. Swapnaja B. More[2]

[12]Aditya College Of Engineering, Maharashtra India

ABSTRACT

Social media allows the creation and interactions of user-created content. Social medium places include Facebook, Twitter etc. Student's casual discussion on social media focused into their educational experience, mind-set, and worry about the learning procedure. Information from such instrumented environments can present valuable data to report student problem. Examining such data can be challenging. The problem of student's experiences reveal from social media content need human analysis. It pays attention on engineering student's Twitter posts to know problem and troubles in their educational practices. This paper proposes a workflow to put together both qualitative investigation and large-scale data mining scheme. First a sample is taken from student and then qualitative analysis conducted on that sample which is associated to engineering student's educational life. It is found that engineering students encounter problems such as heavy learning load, lack of social meeting, and sleep deficiency. Based on this outcome, a multi-label classification algorithm that is Naive Bayes Multi-label Classifier algorithm and Decision tree algorithm is applied to categorize tweets presenting student's problems. The algorithm prepares a detector of student problems. This study presents a tactic and outcome that demonstrate how casual social media data can present insight into student's incident.

Keywords- *Social networking, web-text analysis, Education.*

1. INTRODUCTION

Data mining research has effectively produced several technique, tools, and algorithms for managing huge amounts of data to answer real-world troubles. As social media is widely used for various purposes, vast amounts of user-created data be present and can be made available for data mining. Data mining of social media can enlarge researchers' ability of understanding innovative experience, to the use of social medium and develop business intelligence to present good services and extend innovative opportunities. Main objectives of the data mining procedure are to collectively handle large-scale data, extract actionable patterns, and gain insightful knowledge.. Social media sites such as Twitter, Face book, and YouTube present grand place to students to share happiness and struggle, sentiment and tension, and gain social support. On various social media sites, students talk about their everyday encounters in a comfortable and informal manner. This Student's digital information gives huge amount of implicit information and a whole new viewpoint for educational researchers to know student's experiences outside the prohibited classroom environment. This understanding can enhance education quality, and thus improve student employment, preservation, and achievement [1]. The vast amount of information on social sites provides prospective to recognize student's problem, but it raises some methodological complexities in creating sense of social media data for educational reasons. The complexities such as absolute data volumes, the miscellany of Internet slangs, the change of locations, and moment of students posting on the web. Pure physical analysis cannot contract with the ever growing scale of data, while pure automatic algorithms cannot capture in-depth significance inside the data [2]

The research goal of this learning are:-

1) To show a work of social media information sense making for educational reasons, combining both qualitative investigation and large-scale data mining techniques.
2) To explore engineering student's casual discussions on Twitter, in order to know problem coming into their life.

This Study prefer to focus on engineering student's post comments on Twitter about their problems in collage life because:

1. Engineering schools and branch have long been stressed with student employment and preservation topics. Engineering graduates comprise a significant part of the nation's potential labor force and have a direct impact on the nation's financial expansion [3].

2. Based on understanding of students difficulty decision makers can make more knowledgeable conclusions on proper interference that can help students to conquer obstacles in education.

3. Twitter is a well-liked social media site. Its content is frequently public and very brief that is no more than 140 characters per tweet. Twitter offer free APIs that is used to stream data.

1.1 Social network analysis: Social Networks Analysis (SNA), or structural analysis, aims at studying relationships between individuals, instead of individual attributes or properties. A social network is considered to be a group of people, an organization or social individuals who are connected by social relationships like friendship, cooperative relations, or informative exchange. Different DM techniques have been used to mine social networks in educational environments, but collaborative sorting is the most common. Collaborative filtering or social filtering is a method of making automatic predictions about the interests of a user by collecting taste preferences from many users [3].

1.2 Qualitative analysis: Qualitative analysis is a technique of examination employed in many diverse academic regulation, by tradition in the social sciences, but also in market research and further contexts. Qualitative researchers plan to gather an in-depth understanding of human actions and the reasons that manage such behavior. The qualitative method examines the why and how of decision making, not just what, where, when. Hence, minor but focused samples are often used than huge samples. Qualitative procedures create information only on the particular cases studied, and any more general terminations are only suggestions. Quantitative methods can then be used to look for experimental support for such research theories

2.LITERATURE REVIEW

One of the major research projects regarding engineering student's experiences is the Academic Pathways Study (APS) conducted by the Center for the Advancement of Engineering Education (CAEE). APS consists a series of longitudinal and multi-institutional studies on undergraduate engineering student's learning experiences and their transition to work.. They used various research methods including surveys, structured interviews; semi-structured interviews, engineering design task, and small focus groups. The CAEE website provides 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 [4].

Other smaller research projects usually focus on engineering student's experiences in particular classes. For example, Courter et al. interviewed freshman engineering students about their experiences in a freshman design class using open-ended questions and identified aspects of their experiences that could lead to improved student retention in engineering. Using multiple survey instruments, Demetry and Groccia evaluated and compared mechanical engineering student's experiences in two introductory materials science classes with one implementing active learning and cooperative learning strategies. Torres et al. presented student's experiences of learning robotics within a virtual environment and remote laboratory, where student's knowledge was assessed via automatic correction tests and student's opinions were collected using self-evaluation questionnaires. [5]

"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."[6]. Learning analytics and educational data mining (EDM) 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 educational data, and so, its objective is to analyze these type of data in order to resolve educational research issues.EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners.

2.1 Related Work

The theoretical foundation for the value of informal data on the web can be drawn from Goffman's theory of social performance. Goffman's theory of social performance is widely used to give details on the web today and it is widely used to clarify mediated interactions on the web today. Many studies show that social media users purposefully handle their online identity to "look better" than in real life. Other studies show that there is a lack of awareness about managing online identity among college students, and that young people usually regard social media as their personal space to hang out with peers outside the sight of parents and teachers. Student's online conversations reveal aspects of their experiences that are not easily seen in formal classroom settings, thus are usually not documented in educational literature. Below are reviews of studies on Twitter from the fields of data mining, machine learning, and natural language processing. These studies have more emphasis on statistical models and algorithms. They cover a wide range of topics popularity prediction, event detection, topic discovery and tweet classification. Amongst these topics, tweet classification is most relevant to this study.

Popular classification algorithms include Naive Bayes, Decision Tree, Logistic Regression, Maximum Entropy, Boosting, and Support Vector Machines (SVM).

Most existing studies found on tweet classification are either binary classification on relevant and irrelevant content, or multi-class classification on generic classes such as news, events, opinions, deals, and private messages. Sentiment analysis is another very popular three-class classification on positive, negative, or neutral emotions/opinions. Sentiment analysis is very useful for mining customer opinions on.

Products or companies through their reviews or online posts. It finds wide adoption in marketing and customer relationship management (CRM). Many methods have been developed to mine sentiment from texts our purpose is to achieve deeper and finer understanding of student's experiences especially their learning-related issues and problems. To determine what student problems a tweet indicates is a more complicated task than to determine the sentiment of a tweet even for a human judge. Therefore, our study requires a qualitative analysis.

3. ALGORITHMS USED

This study built a multilevel 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 multilevel classifiers.

3.1 Text Pre-processing

Twitter client use various unusual symbols to express certain significance. For Ex, is used to specify a hashtag, @ used to specify a user account, and RT is used to show a retweet. Twitter users occasionally duplicate letters in words thus to highlight the words, for ex, "soooo cuuuteeee", "verrrryyyy smmmaaaart", and "Looooking awesome". In addition, common stopwords such as "a, an, and, of, he, she, it", nonletter symbols, and punctuation as well carry noise in text. So we preprocessed the texts prior to training the classifier

1) First remove every engineering Problems hashtags. And for new occurring hashtags, just removed the sign, and reserved the hashtag texts.
2) For identifying negative emotion and issues negative words are used. thus it replace words finishing with "n't" and further frequent negative words (e.g. no, not, nothing) as negtoken".
3) Detached every one word that include non-letter symbols and punctuation. This incorporated the deletion of @ and http links. Also delete all the RTs.
4) For replicated letters within words, policy when it discovers two matching letters replicating, it reserved both of them. If it identified more than two same Letters replicating, substitute them with one letter. Therefore, "soooo cuuuteeee" is corrected to "”So cute”. Initially accurate words such as "Sweet" and "buddy" were kept as they were.
5) At this point it used the Lemur data recovery toolkit to eliminate the frequent stop words. It kept words like "much, lot, many, all, forever, still, just", because the tweets regularly use these words to communicate point.

3.2 Naive Bayes Multi-label Classifier

The Naive Bayes classifier is a straightforward probabilistic classifier which is based on Bayes theorem with strong and naive self-government assumptions. It is one of the most basic text categorization method with various applications in email spam exposure, private mail 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. One well-liked way to execute multi-label classifier is to convert the multi-label organization problem into multiple single-label categorization problems [7]

Next is the necessary action of the multi-label Naive Bayes classifier. Assume there are sum of W words in the preparing document compilation in this case, every tweet is a document.

D = d1;d2; : : : ; dw, and a total amount of M categories K = k1; k2; : : : ; kM. If a word dw appears in a category k for

n times, and appear in categories other than k for n ' d $_{w}$k d $_{w}$k
times, then, the probability of this word in a definite category c is

^{n} d $_{w}$**k**

$$p(d_w|k) = \quad \pounds\mathbf{w=i}\ nd_w\mathbf{k}$$

Similarly, the probability of this word in categories other than c is:

$$p^{(d}w|^{k\text{ı})} \quad \underline{{}^{nd}w*} \quad \pounds\mathbf{w=i}\ nd_w\mathbf{k}$$

Suppose there are an entire number of X documents in the preparing set, and K of them are in category k. Then the probability of category k is

$$p(k) = \frac{K}{X'}$$

And the probability of other categories k’ is

$$p(k') = \frac{X - K}{X}$$

For a document di in the trying set, there are Y words Wdi = wi1;wi2; : : : ; wiY, and Wdi is a subset of D. The purpose is to classify this document into category c or not c.

We assume independence among each word in this document, and any word wik conditioned on k or k' follows multinomial distribution. Therefore, according to Bayes Theorem, the probability that di fit in to category k is

$$p(k|di) = {}^{\cancel{P(d}}\mathbf{\cancel{p}}{}^{\Delta\cancel{(k)}a}\mathbf{ny=i}\ p(w_{iy}|k)\ p^{(k)}$$

and the probability that di fit into group other than c is

$$pC^{k'|d}i^{)} = {}^{\cancel{P(d'!}}\text{-}\cancel{p^{f}(d.)}{}^{\cancel{P(k)}}\ {}^{a}\eta\ P\wedge^{Wi}y\,l^{k})\ \ P^{(k<)}$$

Because $p(k|d_;) + p(k' + d_;)$,it normalize the latter two items which are comparative to $p(k|d_;)$ and $pCk'|d_;)$to get the actual values of $p(k|d_;)$. If $p(k|d_;)$is larger than the probability threshold T, then di fit into category k, otherwise, di does fit into category k. Then do again this process for every category. In this execution, if for a definite document, there is none category with a positive probability larger than T, it allocate the one category with the largest probability to this document. In calculation, others is an special group. A tweet is only allocated to others while others is the just category among probability greater than T.

3.3 The tree construction algorithm

This algorithms use a divide and conquer approach to construct a decision tree. It develop a decision tree for a given training set T consisting of set of training instances. An instance indicate values for a rest of attributes and a class. Let the classes be represented by {C1, C2, ..., Cn}. Originally, the class occurrence is computed for instance in training set T. If all instances fit in to similar class, node K with that class is constructed. However, if set T include instances belonging to more than one class, the test for choosing attribute for dividing is carry out and the attribute fulfilling dividing criteria is chosen for the test at the node. The training set T is then partitioned into k restricted subsets {T1, T2, ..., Tk } on the base of this assessment and the algorithm is recursively applied on every nonempty division. The algorithm for creation of a decision tree is given below.

1)Construct (T)

2)Calculate freq (Ci, T).

3)If (all instances belong to same class), return leaf.

4)For every attribute A test for splitting criteria Attribute Satisfying test is test node K.

5)Recur Construct (Ti) on each partition Ti. Add those nodes as children of node K.

6)Stop.

Choosing the best attribute for splitting

The choice of attribute at each node is an main process. Various approaches proposed to choose the best attribute.

Using Information Theory

Objective of process of data categorization is to exploit the information gain as it leads to raise in categorization accuracy.. The gain is describe as the information get from a message based on its probability P. For any set of instances T, the probability that an instance fit in to class Ci is specified as

$$\mathbf{p} = \frac{\text{frequency(Ci,T)}}{|T|} \qquad (1)$$

Where |T| is number of instances in set T and freq (Ci,T) represented the number of instances in T that fit in to class Ci.. Now, the average information enclosed in set T concerning class relationship of instances, called entropy of set T, and is calculated in bits as

$$infor^{(}T^{)} = ZjLi\ Pi\ x\ \log_2 (Pi)bits \qquad (2)$$

k is the number of classes in set T. The test X performed at a node on the chosen attribute gives subsets T1, T2, ., Tk . The information by this dividing process is calculated as the sum over these subsets, is given as

$$infor_x{}^{(}T^{)}\ ZjLi\ \wedge\ x\ infor^{(}T_i{}^{)} \quad (3)$$

The reduction in entropy due to portioning of T with test X on the chosen attribute, represented as Gain (X), is calculated as

$$gain(X) = infor(T) — infor_x(T) \quad (4)$$

The attribute which gives highest information gain is preferred. The difficulty with above approach for collection of test attribute at a node is that it is biased towards attributes with a lot of values as compared to attributes with less values and it leads to large decision trees that weakly generalize the difficulty. This problem can be removed with normalization of gain criterion and utilize of gain ratio. The gain ratio calculates ratio of information generated by portioning T and is expressed as

$$gain\ ratio(X)= gain(X)\ split\ infor(X) \quad (5)$$

The split info(X) calculates information gained by splitting training set T into k subsets on test X.

$$split\ infor(X) = —\ \mathbf{ZjLi}\ ^{x} \log_2 (jp!) \qquad (6)$$

The attribute on which test obtains maximum gain ratio is chosen. This approach has problem that it tends to favour attributes for which split info(X) is very small. Another problem is that gain ratio can be calculated only when the split info(X) is nonzero. To conquer this problem, Quinlan recommended calculate information gain over every one attributes and considering attributes with information gain which is at least as large as average of information gain over all attributes. The use of gain ratio gives improved accuracy and complexity of classifier [8].

4. PROPOSED METHOD

Proposed system developed a workflow to put together both qualitative investigation and large-scale data mining techniques it paying attention on engineering student's Twitter posts to know problem and troubles in their educational practices.. First a sample is taken from student and then it conduct qualitative analysis on that sample which is associated to engineering student's educational life. It found engineering students encounter problems such as heavy learning load, lack of social meeting, and sleep deficiency. Stand on these outcomes, authors apply a multi-label classification algorithm to categorize tweets presenting student's problems. After that used the algorithm to prepare a detector of student problems. This study presents a tactic and outcome that demonstrate how casual social media data can present insight into student's incident.

In this study it implemented a multi-label classification model where we permitted one tweet to go down into many categories at the same time. Our categorization is compared with other generic classifications. Our work expands the range of data-driven approaches in teaching such as learning analytics and educational data mining.

The important point in proposed study are, First, it propose a workflow to bridge and integrate a qualitative research methodology and large scale data mining techniques. It base our data-mining algorithm on qualitative insight resulting from human understanding, so that it can gain deeper understanding of the data. Then apply the algorithm to another large-scale and unexplored dataset, so that the physical method is improved. Second, the paper provides deep insights into engineering student's educational experiences as reacted in informal, uncontrolled environments. Many issues and problems such as study-life balance, lack of sleep, lack of social engagement, and lack of diversity clearly emerge. These could bring awareness to educational researcher, policy- maker.[1]

5.SYSTEM ARCHITECTURE

5.1 System Flow-

The system flow is shown in the figure below: In this system there is an investigative procedure to find the appropriate data and appropriate Twitter hashtags, and then a Twitter hashtag is a word that is starting with a sign, which is used to highlight or tag a issue.

Gathering of tweets using the hashtag engineering- Problems. This corresponds to the step 1 In Fig. 1. Next in the step 2 and 3 of Fig. 1. The inductive content analysis is perform on the sample on engineering problem.

In step 4, it is found that the major problem that comes into engineering students fall into numerous well- known categories. Based on these categories, a multilevel Nave Bayes classification algorithm and decision tree classifier is executed for classification.

In step 5 the performanceof the classifiers is Estimated by comparing it with other multilevel classifiers.

In step 6 the classification algorithm is applied by System to prepare a detector that help recognition of Engineering student's problems. The results are provided by step 7 help educators to identify at risk students and make decisions on proper Interference to preserve them.

Inductive Content Analysis:

There is none predefined categories of the data that is collected, so there is necessity to discover what students saying in the tweets. Thus, it first carry out an inductive content analysis on engineering Problems dataset. Inductive content analysis is one well-likedqualitative research Technique for physically evaluating text content.

Expansion of Categories:

The 5 prominent themes are: heavy study load, lack of social engagement, negative emotion, sleep problems, and diversity issues. Each theme reveals one problem or difficulty that engineering students have in their life. It establishes that many no of tweets fit in to more than one category. For example, "I am fed up of study why I m not in dance school?

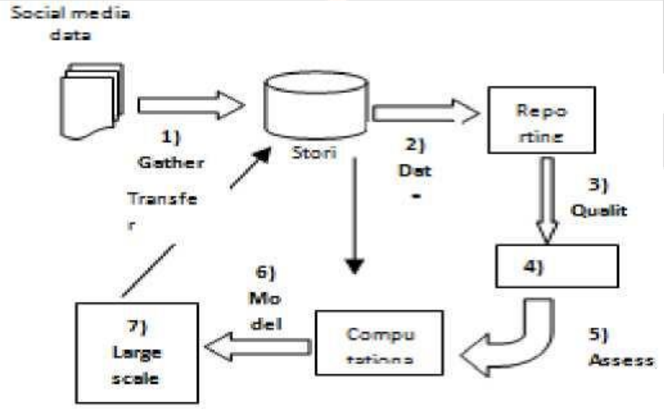

Fig 1: System workflow

Hate being in engineering school. Too much stuff. Way too difficult. No enjoy" comes into heavy study load, and negative emotion at the same time. Hence one tweet can have many categories. This is a multilevel classification as contrast to a single label classification in which each tweet can fall only in one category. The no of categories where one tweet fit in to are called tweets labels otherwise label set.

The Prominent Themes:

If the 1 tweet comes into many categories, it is counted several instances. Amount of tweets in each group examine. Here large amount of tweets fall into "Others". Please note, exemplar tweets presented in each theme may also go down in numerous further categories at the same instance, except ones in "others

6. CONCLUSION

Mining social media data is helpful to researchers in learning analytics, educational data removal, and learning skill. It gives a way to examining social medium statistics that conquer the main restrictions of both physical qualitative analysis and huge scale computational study of user produced textual content. Two algorithms are useful for this classification first the Naive Bayes Multi-label Classifier and Second the Tree construction algorithm. It notifies educational manager, and other applicable assessment makers to expand further accepting of engineering students' institution understanding.

REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan." Mning Social Media Data for Understanding Students' Learning Experiences" IEEE tarnsactions on learning Technologies, ID, DOI 10.n09/TLT.2013.2296520

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

[3] 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.

[4] 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.

[5] R. Ferguson, "The state of learning analytics in 2012: A Review and future challenges," Knowledge Media Institute, Technical Report KMI-2012-01, 2012.

[6] 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.

[7] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-label data,"Data mining and knowledge discovery handbook, pp. 667-685, 2010

[8] Dipak V Patil and R S Bichkar. Article: Issues in Optimization of Decision Tree Learning: A Survey. International Journal of Applied Information Systems 3(5): 13-29, July 2012. Published by Foundation of Computer Science, New York, USA.