IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Data Mining Technique to Understand Students Learning Experience

Meenakshi Das[1], Pinky Saikia Dutta[2]

M.Tech Student,Dept of Computer Science, GIMT, Guwahati, India[1]

Assistant Professor, Dept of Computer Science, GIMT, Guwahati, India[2]

Abstract: Social media sites such as Twitter, Facebook, You-tube are very popular sites in higher educational student's like engineering, medical, pharmacy, trainees and other than students too. It is a platform where one can share their ideas, views and discuss experiences with others in a formal and casual manner. Now a day's gregarious media provides opportunities for understanding human behavior from the large aggregate data sets that their operation collects. Data Mining is very useful in the field of education, especially while examining students' learning behavior. Student's informal discussion on social media (Twitter) elucidates their educational experiences, opinions, feelings, mind-set and concerns about the cognition process. Data from such uninstrumented environments can provide valuableknowledge to apprise student problem. Examining such data, however, can be arduous. The problem of students' experiences reflected from social media content requires human interpretation. However, the growing scale of data demands automatic data analysis techniques. In this paper we propose a workflow to bridge together both qualitative analysis and large-scale data mining techniques. We fixated on engineering students Twitter posts to understand issues and quandaries in their learning experiences. First a sample is taken from student's tweets and then qualitative analysis is conducted on the 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, sleep deficiency etc. Based on this outcome, Naive Bayes Multi-label Classifier algorithm is applied to categorize tweets presenting student's problems. This study presents a methodology and result that demonstrate how casual social media data can provide insight into student's experiences.

Keywords: Data mining, Social networking, Web text analysis, Naïve bayes, Computer and education.

I. INTRODUCTION

Data mining research has produced several techniques, tools, and algorithms for managing huge amounts of data to answer real-world problems. Social media plays important role in today's era of information and technology. As social media is widely used for various purposes, vast amounts of user created datas can be made available for data mining.

Main objectives of the data mining procedure are to communally handle large-scale data, extract useful patterns, and gain useful knowledge. Social media sites such as Twitter, Facebook, and YouTube provides platform for one to share happiness, struggle, sentiment and gain social support. On various social media sites, students talk about their everyday encounters in a comfortable and informal manner.

They share their joys and sorrows related to studies on social media in the form of judgmental comments, tweets, posts etc. Student's digital information gives huge amount of implicit useful and reliable information for educational researchers to understand student's experiences outside the prohibited classroom environment. This understanding can enhance education quality, and thus improve student employment, preservation, and achievement [1][2]. The vast amount of information on social sites provides prospective to recognize student's problem, but also promotes some methodological complexities in use of social media data for educational reasons. Thecomplexities such as assortment of Internet slangs, absolute data

volumes 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 [3].

One important reason why social media can be relayed on is that the comments and posts are spontaneous emotions and feelings of students.

They are not much thought over as we often do while answering any surveys. These studies can be very useful and may prove revolutionary for an educational institute as crucial changes can be made in educational nature of the institute.

The research goal of this learning are:-

1) To make the huge amount of data useful for educational purposes, combining both qualitative analysis and large-scale data mining techniques.
2) To examine engineering students informal tweets on twitter in order to analyze the issues and problems faced by engineering students in their life.

We selected engineering students' problems for our study.

The major reasons were:

1) Engineering schools and branch have long been stressed with student employment and preservation topics. Collages face problems with students' recruitments and retention issues.
2) Engineers (IT industry) comprise a paramount part in growth of GDP of nation and have a direct impact on the nation's financial expansion. So their academic problems must be tackled.
3) Predicated on understanding of student's issues and quandaries, policymakers, educators and difficulty decision makers can make more knowledgeable conclusions on proper interference and services that can help students to conquer obstacles and barriers in education and help the student to solve the problem.
4) Twitter is a well-liked social media site. Its content is often public and very brief, not more than 140 characters per tweet. Twitter provides free APIs that is acclimated to stream data and allows developers to build upon and extend their applications in new and creative ways. To construct a data mining design or to involve in analytics research, the Streaming API is most suitable for such things. Therefore we choose to analyze students post on twitter.

II. LITERATURE REVIEW

One of the main research projects relevant to engineering student's experiences is the Academic Pathways Study (APS) conducted by the Center for the Advancement of Engineering Education (CAEE). APS involves a series of longitudinal and multi-institutional studies on undergraduate engineering student's erudition experiences and their evolution to work. They used various research methods including reviews; interviews, engineering design task, and small focus groups. The CAEE website presents research briefs from the APS study for 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 [12].

Other smaller research projects focuses mainly 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 the open-ended questions and identified aspects of their experiences that could lead to improved student retention in engineering.

Demetry and Groccia, used multiple survey instruments evaluated and then 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 evaluated via automatic correction tests and student's opinions were collected using self-evaluation questionnaires [13].

Related Work

The theoretical basis for the value of informal data on the web can be drawn from Goffman's theory of social performance [4]. Goffman's theory of social performance is widely used to give detail of mediated interactions on the web today [5] .One of the most fundamental aspects of this theory is the notion of front-stage and back-stage of people's social performances. Compared with the frontstage, the relaxing atmosphere of back-stage usually applauds more spontaneous actions. For students, compared to formalclassroom settings, social media are relative informal and relaxing back-stage. When students post content on social media sites, they usually post what they think and feel at that moment. In this sense, the data collected from online conversation may be more authentic and unfiltered than responses to formal research prompts. Many studies show that social media users may purposefully manage their online identity to "look better "than in real life [6] [7] .Other studies show that there is dearth of awareness about managing online identity among college students [8], and that young people usually regard social media as their personal space to hang out with peers outside the sight of parents and teachers [9]. Students' online conversations reveal aspects of their experiences that are not easily seen in formal classroom settings, thus are usually not documented in educational literature. The abundance of social media data provides opportunities but also presents methodological difficulties for analyzing large-scale informal textual data. The next section reviews popular methods used for analyzing Twitter data.

Gaffney [10] analyses tweets with hashtag #iranElection using histograms, user networks, and frequencies of top keywords to measure online activism. Similar studies have been conducted in other fields including healthcare [12], marketing [13], and athletics [14]. 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. 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).

III. MATHEMATICAL BACKGROUND OF OUR WORK

The Naive Bayes classifier is an open probabilistic classifier which is based on Bayes theorem. Naive Bayes executes well in many difficult real-world troubles. Naive Bayes classifier is extremely efficient since it is less computational 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 [15]. Following are the basic procedures of the multi-label Naive Bayes classifier.

Let there are sum of N words in the training document collection (every tweet is a document). W $=w_{1,}w_{2,}$

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

…$w_{N,}$and a total number of L categories C = $c_{1,}c_{2}$, …$c_{L.}$ If a word w_n appears in a category c form$_{wnc}$ times, and appear in categories other than c for $m_{wnc'}$ times. Then, the probability of this word in a definite category c is

$$p(w_n|c)=\frac{m_{wnc}}{\sum_{n=1}^{N} m_{wnc}}$$

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

$$p(w_n|c')=\frac{m_{wnc'}}{\sum_{n=1}^{N} m_{wnc'}},$$

Suppose there are M documents in the training set, and C of them are in category c. Then the probability of category c is

$$p(c) = \frac{C}{M}$$

And the probability of other categories c' is

$$p(c') = \frac{M-C}{M}$$

For a document d_i in the testing set, there are K words, W_{di}= $w_{i1}, w_{i2}, \ldots . w_{iK}$, and W_{di} is a subset of W.Our purpose is to classify this document into category c or not c. We assume independence among each word in this document, and any word w_{ik} conditioned on c or c' follows multinomial distribution. Therefore, according to Bayes Theorem, the probability that d_i fit in to category c is

$$p(c|d_i) = \frac{p(d_i|c).p(c)}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c).p(c)$$

and the probability that $d_i$fit into group other than c is

$$p(c'|d_i) = \frac{p(d_i|c').p(c')}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c').p(c')$$

Because $p(c|d_i) + p(c'|d_i) = 1$, we normalize the latter two items which are comparative to $p(c|d_i)$ and $p(c'|d_i)$ to get the actual values of $p(c|d_i)$. If $p(c|d_i)$is larger than the probability threshold T, then $d_i$fit into category c, otherwise, d_i does fit into category c. Then do this process again 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 a special group. A tweet is only allocated to others while others is the just category among probability greater than T.

IV. OUR APPROACH

We develop a workflow to put together both qualitative investigation and large-scale data mining techniques. Primarily a sample is taken of students tweets related to engineering students learning life and then inductive content analysis on that sample.

It found engineering students encounter problems such as heavy study loads, lack of social engagement, sleep problem etc.

Based on these outcomes, we apply a multi-label classification algorithm to categorize tweets presenting student's problems. This study presents a tactic and outcome that demonstrate how casual social media data can present insight into student's incident.

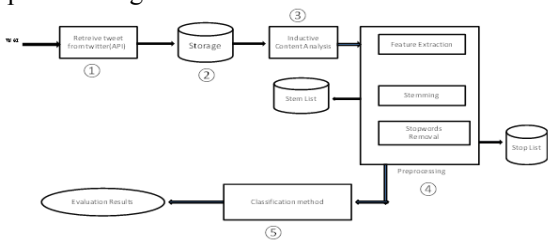

Fig 1: System Workflow

IV. I. SYTEM ARCHITECTURE

Datas from the students post on twitter are collected in database; extraction is done on the data. Results are then collected and noisy data is removed. Refining the data from the store gives the model training evaluation. Finally by the model adaption, large scale data analysis result is generated. In the step 1, we collect posts by engineering students on the social networking site. The tweets are stored in database and inductive content analysis is then performed on the database in step 2 and 3.We categorize their several problems in prominent categorizes. Pre-processing of the data's are done in step 4.Based on these categories, we are implementing a multi-label Naïve Bayes classification algorithm in step 5.Thus the evaluation results could help educators to identify at-risk students and make decisions on proper interventions to retain them.

IV.II. DATA COLLECTION

We are collecting all the information from the different student's posts in social website twitter. Twitter's tweets have been used as text data. Twitter data that is publically available were collected by Twitter API. Twitter's streaming API service is used to store real-time tweets. Tweets will be collected from the account by using the secret tokens of the twitter application. Twitter authenticates the secret tokens and allows the user to access the twitter to collects the tweets. Twitter data that is publically available were collected by Twitter API. Python's API named Tweepy [11] have been used to implement streaming API of Twitter. It provides libraries to collect streaming twitter data. The incoming tweets were stored in CSV (Comma Separated Values) file format in real-time by importing Python's CSV library functions.

IV.III. INDUCTIVE CONTENT ANALYSIS

Because social media content like tweets contains large amount of informal languages that cannot be used for analysis is often ambiguous and subject to human interpretation. The unsupervised algorithm does not detect the informal conversations; it need proper training for analysis of data. The data are not specified in any category so we needed to explore what students were revealing in the tweets. Thus, first need to conduct an inductive content

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

analysis on the #engineering Problems data set. The Inductive content analysis is one popular qualitative research method for manually analyzing text content.

IV. DEVELOPMENT OF CATEGORIES

Various types of emotions related to learning experiences are termed as prominent themes. Below are the few themes related to the project:

1) Heavy Study Load: Students who face problems which are dominated by labs, exams, homework and classes will be grouped to heavy study load category.

2) Lack of Social Engagement: Students feel that they give up their freedom and joy time for the sake of the academic works.

3) Negative Emotion: Expressing the anger, sickness, depression and disappointment will come under negative emotions.

4) Sleep Problems: Students frequently suffer from lack of sleep and nightmares due to heavy study load and stress.

5) Diversity Issues: Anti-social image of mingling with others is coined as the diversity issue. Without knowing the background all the negative comments will passed on which may cause the diversity issue.

6) Positive Emotion: Not all tweet about the difficulties faced in their learning courses. But sometimes also about the good things of getting jobs, good marks, happiness.

7) Others: A large number of tweets fall under this category. Many tweets in this category do not have a clear meaning. Other tweets in this category do reflect various issues that engineering students have but seems in small volumes.

Examples are curriculum problems, lack of motivation, career and future worries identity crisis and physical health problems

IV.V. TEXT PRE-PROCESSING

Data pre-processing reduces the size of the input text significantly. It encompasses activities like sentence boundary determination, natural language specific stop-word elimination and stemming. Stop-words are handy words that occur frequently in the language of the text (for example, “a”, “the”, “an”, “of” etc. in English language), this are not useful for classification. Stemming is reducing of words to their root or base form.

For English language, the Porter’s stemmer is a popular algorithm, which is a suffix stripping sequence of systematic steps for stemming. Stemming reduces the vocabulary of the training text by approximately one-third of its original size.

For example, using the Porter’s stemmer, the English word “generalizations” subsequently be stemmed as “generalizations → generalization → generalize → general → genre”.

1) Feature Extraction

The feature extraction process is derived from Bag-of words [16] approach. (Here the text is represented as a bag of its words). The frequency of occurrence of each word is used as a feature for training the classifier Naïve Bayes algorithm.

2) Removal of Stopwords

A word that occurs in 80% of the documents in the collection is abortive for purposes of retrieval. Such words are referred to as stopwords and are mostly filtered out as possible index terms. Articles, prepositions, and conjunctions are natural contenders for a list of stopwords. Removal of stop words has additional important benefit. It reduces the size of the indexing structure broadly. In fact, it is typical to obtain a compression in the size of the indexing structure of 40% or more solely with the elimination of stopwords. Despite these benefit, elimination of stopwords might reduce recall. For example, a documents containing the phrase “to be or not to be.” Elimination of stopwords might leave only the term be making it almost impossible to properly recognize the documents which contain the phrase specified.

3) Stemming

Mostly, the user specifies a word in a database but only a variant of this word is present in a relevant document. This can be partially overcome with the substitution of the words by their respective stems. A stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes). Stems are useful for improving retrieval performance because they reduce variants of the same root word to a common concept. Moreover, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced.

So we pre-processed the texts before training the classifier.

1) We remove all the #engineering Problems hashtags. For other co-occurring hashtags, we only removed the #Sign, and keep the hashtag texts.
2) Negative words are helpful for detecting negative emotion and issues. Therefore we substituted words ending with “n’t” and other common negative words (e.g., nothing, never, none, cannot) with “negtoken”.
3) We remove all words that contain non-letter symbols and punctuation. This includes removal of @ and http links. We also removed all the RTs.
4) For repeating letters in words, when we detected two identical letters repeating, we kept both of them. If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, “huuungryyy” and “sooo” were corrected to “hungry” and “so”. “muuchh” was kept as “muuchh”.

IV.VI. NAÏVE BAYES MULTI LABEL CLASSIFIER

We built a multi-label classifier to classify tweets based on the categories developed in the previous content analysis stage. There are several popular classifiers widely used in data mining and machine learning domain. Out of them we

DOI 10.17148/IJARCCE.2016.5582

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

use naive bayes algorithm. We first calculate the probability of particular word belongs to specific category C and belongs to category other than C.

It is calculated as:-

$$p(c|d_i) = \frac{p(d_i|c).p(c)}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c).p(c)$$

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

$$p(c'|d_i) = \frac{p(d_i|c').p(c')}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c').p(c')$$

In our study the accuracy of the model has been found out by the confusion matrix by the formula given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

V. EXPERIMENTAL RESULT

Confusion Matrix

Classified as	->	a	b	c	d	e	f	g
Diversity Issues	a	<32>	3	.	5	18	4	.
Heavy Study Load	b	.	<128>	.	.	3	1	.
Lack Of Social Engagement	c	.	15	<22>	5	8	3	.
Negative Emotion	d	.	4	.	<120>	8	.	.
Others	e	.	.	.	.	<160>	.	.
Positive Emotion	f	.	1	.	3	13	<97>	.
Sleep Problems	g	.	8	.	5	10	1	<37>

Accuracy =83.35664335664336

.Table.1 Confusion Matrix

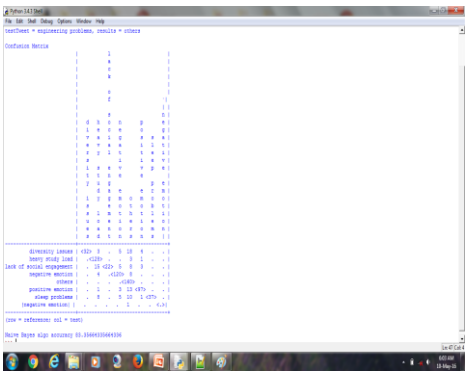

Fig. 2 Screenshot of evaluation result

VI. CONCLUSIONS

Mining social media data is benign for researcher in an education system to detect students learning experience. The workflow for analyzing educational content in the social media helps to overcome the drawbacks of large data mining and manual quality analysis of user generated textual content in social media. We have assimilated qualitative analysis along with data mining techniques. The descriptive process classifies engineering students' twitter data into their belonging problems and perks. The descriptive categories guide the future students in their selection of educational stream. Moreover, based on the categories, the quality of education system can be improved further. So, in our study a multi label classification model is used which will make fall tweet into multi-categories at same time. This work helps the organization and an educational system to understand the problems present during student educational experience. Based on this organization and participation can easily take decision in the engineering studies. Further we can enhance in image processing like images, emoticons, videos etc. This system is also benefial 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

[1]. Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan "Mining Social Media Data for Understanding Students' Learning Experiences," IEEE transactions on learning technologies, manuscript id 1, pp. 1-14, 2013

[2]. Pallavi K. Pagare, "Analyzing Social Media Data for Understanding Student's Problem" International Journal of Computer Applications (0975 – 8887), 2014.

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

[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]. Ajay Kumar Pal, Saurabh Pal, "Classification model of Predication for Placement of Students," in I.J. modern Education and Computer Science, pp. 49-56, November 2013.

[6]. Hsin-Ying Wu, Kuan-Liang Liu and Charles Trappey, "Understanding Customers Using Facebook Pages: Data Mining Users Feedback Using Text Analysis", IEEE 18th International Conference on Computer Supported Cooperative Work in Design, 2014, pp. 346-350..

[7]. I-Hsien Ting, Shyue-Liang Wang, Hsing-Miao Chi and Jyun-Sing Wu, "Content Matters: A study of hate groups detection based on social networks analysis and web mining",IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013,pp. 1196-1201.

[8]. Sara Keretna, Ahmad Hossny and Doug Creighton," Recognising User Identity in Twitter Social Networks via Text Mining", IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 3079-3082

[9]. PAPER 5: Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary, "Twitter Trending Topic Classification", 11th IEEE International Conference on Data Mining Workshops, 2011 pp. 251-258.

[10]. N. Anitha, B. Anitha, S. Pradeepa, "Sentiment Classification Approaches – A Review," in International Journal of Innovations in Engineering and Technology, vol. 3, Issue 1, pp. 22-31, October 2013

[11]. "Using the Twitter Search API"|Twitter Developers, https://dev.twitter.com/docs/using-search, 2013.

[12]. CAEE Academic Pathways Study. at www.engr.washington.edu/caee/about_APS.html

[13]. Atman, C. et al. Enabling engineering student success: The final report for the Center for the Advancement ofEngineering Education. (2010).

[14]. Chachra, D., Kilgore, D. & Loshbaugh, H. G. Being and becoming: gender and identity formation of engineeringstudents. American Society for Engineering Education Annual Conference (2008).

[15]. Robotics at the University of Alicante. International Journal of Engineering Education **22**, 766-776 (2006).

[16]. Bag-of-Words feature extraction technique: https://en.wikipedia.org/wiki/Bag-of-words_model