Classification of Social Media Statistics for Students Perceptive in Learning by ROST

Lakshmi Priya. N,
PG-Information Technology,
Jayam college of Engg. and Tech.
Dharmapuri,India.

Saravanakumar. R
Assistant Professor,
Jayam college of Engg. and Tech.
Dharmapuri,India.

Sivakumar. C
Associate Professor
Jayam college of Engg and Tech.
Dharmapuri,India.

***Abstract*—Social media services provides the great opportunity for students to discuss and divide their everyday meet problems and issues in an easy and casual manner. Such data can be very challenging for analysis. The convolution of students' understanding emulated from social media content requires anthropological awareness. However, the developing ratio of data requirement computed data analysis techniques. It mainly focused for enlightening purpose to describe the process of social media information sense-making and to incorporate both approximate analysis and significant information mining techniques and to examine engineering students' familiar conversations on social media, in order to penetrate the unease and trouble students meet in their learning experiences and this also gives the preservation of students. In preceding system the Naive Bayes classifier is used to classify the twitter dataset to increases the mean squared error. The ROS Tree classification technique use to prove the proposed works effectiveness**.

Keywords*—*Classification techniques, Naive Bayes Multi-Label, Social media information, Social media services, ROST.

I. INTRODUCTION

Students converse and divided their day by day encounters difficulties and issues in an easy and contributory method on social media sites. Students digital tread provides enormous amount of implied knowledge and new aspective for educational researchers and learners to realize students experiences from outside classroom environment. This learning can enlighten institutional management on interventions for at threat students, enhancement of learning feature, and thus augment the student employment, preservation, and accomplishment.

The rising pasture of education analytics and enlightening statistics mining has persistent on analyzing prepared data obtained from route organization systems (ROS), classroom knowledge convention, or inhibited online education environments to enlighten learning management. To the best of our edification, there is no examination establish to directly extract and analyze which is student- posted content from unconstrained spaces on the social network with the clear aspiration of accepting students' education experiences.

A. SOCIAL MEDIA INFORMATION

Social media information is quite diverse from the conventional data that we are recognizable with in data mining. Apart from huge size, the mainly user-generated data is noisy and formless, with abundant social associations such as friendships and groups. The great demand for recent techniques ushers in and entails a new interdisciplinary field social media mining.

Social media is the social interaction among people in which they produce, dividend or transfer information and design in virtual communities and networks. Social media is distinct as a collection of information-based applications that construct on the ideological and technical information of Web 2.0 and that permit the design and interactions of user-generated substance. Social media is firm of diverse types of social media sites as well as conventional media such as newspaper, radio, and small screen and non-conventional media such as Face book, Twitter, etc.

The examiner objective of this study are 1) For educational purpose to develop a workflow of social media information sense-making and to integrate both qualitative analysis and significant data mining techniques 2) to recognize students issues and troubles in their education experience, in order to discover engineering students' familiar conversations on Twitter.

This study is beneficial to researcher in the field of learning analytics and educational data mining and enlightening technologies.

II. LITERATURE REVIEW

Twitter is classifying the sentiment twitter communication on movie reviews by using machine learning algorithm and is used feature removal method for automatically mine the messages.Classification will only work on tweets in English because the training data is English-only. But it should be possible for other language sentences. It has limit of tweets in response for any request. It focused domain of movie reviews, blogs and product reviews.There is no appropriate unsupervised algorithm could expose importance in these data. Latent Dirichlet Allocation (LDA) issue the modeling algorithm. It has only produced meaningless word groups from such that data with lot of overlapping words across different topics [2].

Rost et al [5] 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. To concur with this argument, as found no appropriate unsupervised algorithms could reveal in-depth meanings in our data. For example, LDA (Latent Dirichlet Allocation) is a popular topic modelling 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.In existing many classification algorithm used such as SVM , ID3 and Naïve Bayes Classification to classify the tweets on social media data.

III. PROPOSED TECHNIQUE

A. *DATA COLLECTION*

The tool namely called as Radian6 was used to collect the information related to students learning experiences. It can be very challenging to search data because they are different types of language used. Searching data on educational account based on the Boolean combination of keywords like engineer, students, campus, class etc.

In Fig, first step is to collect the data which is student generated on the twitter using radian 6 after that take the random sample of tweets from the data collection for researcher conduct an inductive substance analysis for that data then provide the feedback and description of the codebook for student affected the major problem and send it to other Researcher for review. Two researchers analysed then come to one solution for the above problem. So based on the researcher evaluations send the qualitative result after that using Reduced Offense Skive Tree for efficient classification. Finally give the efficient result.

Social media content like tweets contain a huge quantity of informal language, acronyms, and misspellings, significance is frequently uncertain and subject to being analysis. There were no pre-defined categories of the data, so required to discover what students be saying in the tweets and have to conduct

an inductive substance testing on the #engineering Problems dataset. Inductive substance testing is one of the well-liked qualitative research system for manually analyzing the text substance. Three researchers accomplish the substance analysis process. Analysis was to identify what are the major problems, concerns, and issues that engineering students meet in their learning and existence.

Researcher A read a random sample of tweets from the distinctive #engineering Problems tweets, and developed 13 initial categories including: curriculum problems, heavy study load, study difficulties, imbalanced life, future and carrier worries, lack of gender diversity, sleep problems,

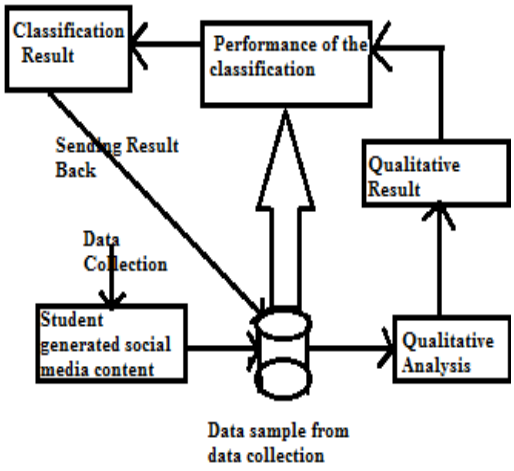

Fig.1 Workflow Diagram for Qualitative Data and Analysis

stress, lack of motivation, physical health problems, nerdy culture, identity crisis, and others.These were developed to categorize as many issues as possible, with no accounting for their virtual significances. Researcher A wrote complete descriptions and gave examples for all category and sent the codebook and the 2,000-tweet sample to researchers B and C for review.

Then, the three researchers discussed and distorted the initial categories into five major themes, because they were themes with quite large number of tweets. The five major themes are: heavy study load, lack of social engagement, negative emotion, sleep problems, and diversity issues. Each theme reflects one issue or problem that engineering students meet in their learning experiences.

In this examiner, create that many tweets could belong to more than one category. For example, “This could very well turn into an all nighters...f*** you lab report #no sleep” falls into heavy study load, lack of sleep, and negative emotion at the same time. “Why am I not in business school?? Hate being in Engineering school. Way too complicated. No fun” falls into heavy study load, and negative emotion at the same time. So one tweet can be present with several categories. This is a multi-label classification as conflicting to a single-label classification where each tweet can only be labeled with one category. The categories one tweet belongs to are called this tweet’s labels or label set.

B. *INTER RATER AGGREMENT*

Statistical measures are commonly used to report concurrence between researchers in substance analysis fiction. However, these methods can only be used for data that belongs to mutual categories. Because we were trading with a multi-label classification problem with non-mutual categories. We used F_1 measure which is used to find the harmonic mean between two sets of data. F_1 Score is 1 if the two sets of data are closely the identical, and is 0 when the two sets of data are completely dissimilar. It represents how close two label sets are assigned to one tweet by two researchers.

Then calculated the F_1 scores among the label sets known by any two researchers to a tweet, and then averaged in excess of all the tweets to represent the concurrence. Assume if there are a total number of N tweets can be categorized by two researchers for the i[th] tweet, x_{1i} represents the amount of labels specified to this tweet by researcher A, x_{2i} indicates the quantity of labels given to this tweet by researcher B, and s_i identify the number of labels that are familiar among researcher A and researcher B (the agreed number of labels).Let $p_{1i} = {}^{s_i}/_{x_{1i}}$ and $p_{2i} = {}^{s_i}/_{x_{2i}}$ then

$$F_1 = \frac{1}{N}\sum_{i-1}^{N}\frac{2p_{1i}\ .p_{2i}}{p_{1i}+p_{2i}}$$

After we decided on the six categories in this study (heavy study load, lack of social engagement, negative emotion, sleep problems, diversity issues and others), took a arbitrary model of 500 tweets, and characterized them individually. If a tweet does not express any of the five major problems, it is considered as "others". A tweet in "others" can be an engineering student difficulty other than the five major ones, or a noisy tweet that does not have understandable significance. Unlike the five major themes, "others" is a restricted category.

The F_1 scores among any of the two researchers were $F_{1AB} = 0.7972$, and $F_{1BC} = 0.7859$and $F_{1AC} = 0.8179$ .The three researchers then discussed tweets were discarded, and we analysed a further arbitrary model of 500 tweets. The F_1scores then augmented to $F_{1AB} = 0.8104$, and $F_{1BC} = 0.8011$ and $F_{1AC} = 0.8252$ correspondingly. For the second 500 tweet example, we used only the normally approved labels by the three researchers for every tweet in our presently computation phase .If there was no connection between the three researchers label sets for a assured tweet, the tweet was unused. Exposed of the 500 tweets, 405 were reserved and used in model education and testing. Researcher A then completed analysing a further arbitrary model of 2,380 tweets. Plus the 405 tweets, there were a total of 2,785 labeled tweets used for reproduction education and testing.

C. REDUCED OFFENSE SKIVE TREE

Reduced Offense Skive Tree (ROST) is fast decision tree learning reduces the size of decision tree by removing sections of the tree that offer little power to categorize instances. The objective of skiving is reduced density of the ending classifier as well as better predictive correctness by the reduction of overfitting and exclusion of sections of a classifier that may be based on noisy or erroneous data. The ROS methods ensure for all central node, whether replacing it with the most recurrent class that does not reduce the correctness of trees. In this case, the node is skived. The procedure continues in anticipation of any further skiving would decline the accuracy. Skiving put in C4.5 by replacing the central node with a leaf node in that way reducing the error rate. ID3 and C4.5 accepts both continuous and categorical attributes in generating the decision tree. It has an improved method of tree skiving that reduces mistaxonomy errors due to noise or in large amounts details in the training Data set.

A decision tree with binary splits for regression. An objective of class Regression Tree can predict responses for new data with the predict method. The object contains the data used for training, so can compute reconstitution predictions. Regression Tree RT(x, y) proceeds a regression tree based on the input variables x and output (response) y. Regression Tree is a binary tree where each branching node is divide based on the values of a column of x.

Discovering a binary question which gives the greatest information about the Y should be recognized and the method should replicate for all levels of the tree. Here Y is measured to be the spam branch of the tree. The leaf X should provide the maximum information about this branch that better discriminates the spam and valid sites. In each child node process should be frequent in greedy method. And finally it yields a tree with maximum information gain of spam websites. Since the algorithm is recursive it requires stopping criteria. It is a threshold here. The sum of squared errors for a tree RT is defined as,

$$S = \sum_{c\in leaves(T)}{}' \sum_{i\in C}{}' (y_i - m_c)^2$$

Where $m_c = \frac{1}{n_c}\sum_{i\in C} y_i$,the calculation for leaf c. The above formula can be written as

$$S = \sum_{c\in leaves(T)}{}' n_c v_c$$

Where, $n_c$is considered to be the within variance and $v_c$is the class prediction

IV. IMPLEMENTATION

To evaluate the performance of classification models commonly used measures include accuracy, precision, recall, and the harmonic average between precision and recall the F_1 score. For multi-label classification, the situation is somewhat more difficult, because each document gets assigned multiple labels. With these labels, some may be accurate, and others may be erroneous. Therefore, there are usually two types of evaluation measures.

a. Example based measures and Label-based measures.

Example based measures are calculate on each document (e.g. each tweet is a document, and also called an example) and then averaged in excess of all documents in the dataset, whereas label-based measure are calculated based o each label(category) and then averaged over all labels(categories).

Accuracy

$$Accuracy\ a = \frac{tp+tn}{tp+tn+fp+fn}$$

Where,

tp - True positive is correctly identified
fp - False positive is incorrectly identified
tn - True negative is correctly rejected
fn -False negative is incorrectly rejected

V. CLASSIFICATION RESULT

From the inductive substance analysis stage, we had a total of 2,785 #engineering Problems tweets annotated with 6 categories. We used 70% of the 2,785 tweets for training (1,950 tweets), and 30% for testing (835 tweets). 85.5% (532/622) of words occurred more than once in the testing sets were found in the training data set. Table 1 shows the 6 evaluation measures at each probability threshold values from 0 to 1 with a segment of 0.1.We assigned the one category with the largest probability value to the document when there was no category with a positive probability value larger than T. So when the probability threshold was 1, it was equivalent to outputting the largest possible one category for all the tweets.

With five multi-label categories and one "others" category, there are (25-1) +1=32 possible label sets for a tweet. Table 1 and Table 2 present all the evaluation measures below arbitrary guessing. The arbitrary guessing program first guessed whether a tweet belongs to "others" based on the proportion this category takes in the training dataset. If this tweet did not belong to "others", it then proceeded to guess whether it fell into the rest of the categories also based the proportion each category takes in the rest categories.

TABLE I

Evaluation Measures with SVM Classifier under Different Probability Thresholds

Probability Threshold	**Example-based accuracy**	**Example-based precision**	**Micro-averaged F1**	**Macro-averaged F1**
0	0.1720	0.1720	0.2940	0.2550
0.2	0.6620	0.6670	0.6840	0.6795
0.5	0.7018	0.7090	0.7050	0.6115
0.8	0.7060	0.7156	0.7050	0.6005
1	0.7086	0.7198	0.7076	0.6026
Rand	0.0414	0.0416	0.0392	0.0180
0	0.1720	0.1720	0.2940	0.2550

From Table 1, see that when the probability threshold value is 0.4, the performance is generally better than under other threshold values.

With five multi-label categories and one "others" category, there are (25-1) +1=32 possible label sets for a tweet. Table 1 and Table 2 present all the evaluations measures below arbitrary

TABLE II

Evaluation Measures

Example-based accuracy	**Example-based precision**	**Example-based recall**	**Example-based F1**
0.5518	0.6058	0.5830	0.5800

guessing. The arbitrary guessing program first guessed whether a tweet belongs to "others" based on the proportion this category takes in the training dataset. If this tweet did not belong to "others", it then proceeded to guess whether it fell into the rest of the categories also based the proportion each category takes in the rest categories.

VI. CONCLUSION

The study is useful to understand the physical education phenomenon that occurs in the field of learning analytics, educational data mining, and enlightening technologies. Through this study 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 too many consequences including lack of social engagement, sleep problems, and other psychological and physical health problems.

Many students feel engineering is uninteresting and inflexible, which leads to lack of motivation to study and negative emotions. Diversity issues also reveal culture conflicts and culture stereotypes existing among engineering students. Building on top of the qualitative insights, now implemented and evaluated a multi-label classifier to detect engineering student problems from University.

This detector can be applied as a monitoring mechanism to identify vulnerable students at a specific university in the long run without repeating the manual work frequently. This is beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and significant computational analysis of user-generated textual content. This study is beneficial to researcher in the field of learning analytics and educational data mining and enlightening technologies.

REFERENCES

[1] Arnold. K.E and Pistilli.M.D, "Course signals at Purdue: Using learning analytics to increase student success," in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp.267-270.

[2] Baker.R and Yacef. K, "The state of educational data mining in 2009: A review and future vision," Journal of Educational Data Mining, vol1, no.1, pp.3-17, 2009.

[3] Becker.H, Naaman.M, and Gravano.L, "Selecting quality Twitter content for events," in Proceedings of the 5th international AAAI conference on Weblogs and Social Media 2011.

[4] Cetintas.S, Si.L, Aagard.H, Bowen.K, and M. Cordova- Sanchez, "Microblogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging- Supported Classroom," Learning Technologies, IEEE Transactions on, vol. 4, no. 4, pp. 292–300, 2011.

[5] Yang.J and Counts.S, "Predicting the speed, scale, and range of information diffusion in twitter," Proc. ICWSM, 2010.