Text Categorization using Data Mining Technique on Social Media Data

[I]Remya R S, [II]Smitha E S

[I]Kerala University, LBSITW, Poojappura, Thiruvananthapuram, India

[II]Dept. of Computer Science, Kerala University, LBSITW, Poojappura,Thiruvananthapuram, India

Abstract

Social media provide a range of opportunities for understanding human behavior through the large aggregate data sets that their operation collects. Data Mining is very useful in the field of education especially when examining students learning behavior in online learning environment. Students casual conversations on social media show their educational experiences, opinions, feelings and concern about the learning process. The data extracted from the social media provide valuable knowledge to inform student learning experience about how they positively taking the education and what are the benefits they gain from the education and what are their problems and issues. Existing methods are usually very time consuming and scale of such studies is also limited. This study focuses on engineering students to understand issues and problems in their educational experiences. Usually engineering students encounter problems such as heavy study load, lack of social engagement, and sleep deprivation. Based on these results, a multi-label classification algorithm is used to classify comments in which attributes are divided into multiple categories reflecting students problems. This understanding can inform institutional decision-making on interventions for at-risk students, improvement of education quality, and thus enhance student recruitment, retention, and success.

Keywords

Data Mining, Learning Analytics And Knowledge, Educational Data Mining, Sentiment Analysis.

I. Introduction

Social Networking Internet services are changing the way to communicate with others and entertain. The rapid growth in popularity of social networks has enabled large numbers of users to communicate, create and share content, give and receive recommendations. Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Social media provide a range of opportunities for understanding human behavior through the large aggregate data sets that their operation collects. Data mining can be defined as the process involved in extracting interesting, interpretable, useful and novel information from data. The objective of data mining in each application area is different.

Data Mining is very useful in the field of education especially when examining students learning behavior in online learning environment. 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, Facebook, and YouTube provide great venues for students to share joy and struggle vent emotion and stress and seek social support. On various social media sites, students discuss and share their everyday encounters in an informal and casual manner. The data extracted from the social media provide valuable knowledge to inform student learning experience about how they positively taking the education and what are the benefits they gain from the education and what are their problems and issues. Academic problems of students comes in various forms such as difficulty in maths subject, lack of motivation, study habits, strict teachers and failed in major examinations [7]. Identifying these problems along with their negative attitude towards education can inform institutional decision-making on interventions for at-risk students, improvement of education quality, and thus enhance student recruitment, retention, and success.

II. Literature Survey

Traditionally, educational researchers have been using methods such as surveys, interviews, various classroom activities to collect data related to students learning experiences. Usually these methods are very time consuming, thus cannot be duplicated or repeated with high frequency. And the scale of such studies is usually limited.

For instance, the Academic Pathways Study (APS) is a five year, multi-institution study. The APS is the largest element of the Center for the Advancement of Engineering Education (CAEE). The primary goal of the Academic Pathways Study was to create a rich and wide-ranging portrait of the undergraduate engineering learning experience. It uses multiple research methods and rely on the students own words for much of the data.

A. Learning Analytics and Knowledge

Learning analytics and knowledge is the measurement, collection, analysis and reporting of data about learners and their contexts for the purpose of understanding and optimizing learning and environments in which it occurs [3]. It is used for prediction purposes, personalization and adaptation and intervention purposes.

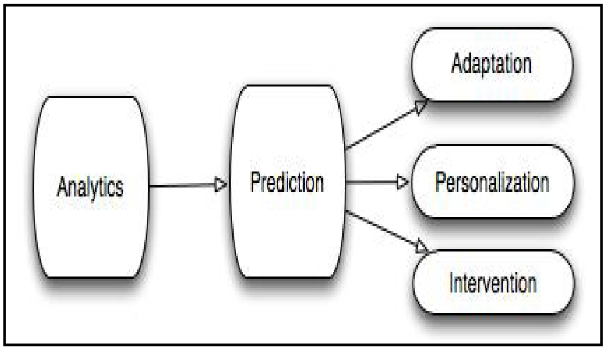

Fig.1: learning analytics

It is used to access academic progress, predict future performance and spot potential issues. Data are collected from explicit student actions and tacit actions. Explicit student actions include completing assignments and taking exams. And tacit actions include online social interactions, extracurricular activities and other activities that are not directly assessed as part of the student's educational progress. The goal of learning analytics is to enable teachers to tailor educational opportunities to each student's level of need and ability.

Learning analytics and knowledge includes the following:

- Social network analysis (e.g., analysis of student-to-student and student-to-teacher relationships and interactions to identify disconnected students)
- Social or attention metadata to determine what a user is engaged with.

B. Educational Data Mining

Data mining can be defined as the process involved in extracting interesting, interpretable, useful and novel information from data. The objective of data mining in each application area is different. Educational Data Mining is the application of data mining techniques to educational data [2]. Its objective is to resolve educational research issues. The first phase of the EDM process is discovering relationships in data. Discovered relationships must then be validated inorder to avoid overfitting. Validated relationships are applied to make predictions about future events in the learning environment. Predictions are used to support decision-making processes and policy decisions.

C. Sentiment Analysis

Social Networking Internet services are changing the way to communicate with others and entertain. The rapid growth in popularity of social networks has enabled large numbers of users to communicate, create and share content, give and receive recommendations. Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Social media sites such as Twitter, Facebook, MySpace and YouTube provide great venues for students to share their happiness, stress and seek social support. On various social media sites, students discuss and share their everyday encounters in an informal and casual manner. Most existing studies found on social media data classification are either binary classification or multi-class classification.

Sentiment analysis is also known as opinion mining. It is a three-class classification on positive, negative, or neutral opinions [6]. Sentiment analysis is very useful for mining customer opinions on products. It can help marketers evaluate the success of an ad campaign or new product launch, determine which versions of a product or service are popular and identify which demographics like or dislike particular product features. Only knowing the sentiment of student-posted comments in social media does not provide much knowledge on relevant interventions and services for students. The purpose is to achieve deeper and finer understanding of student's experiences especially their learning-related issues and problems.

III. Proposed System

The role of this work is to demonstrate the role played by social media for educational purposes and to explore student's informal conversation on social media sites, inorder to understand their issues and problems. A Multi-label classifier is implemented to classify the students posted contents collected from the social media.

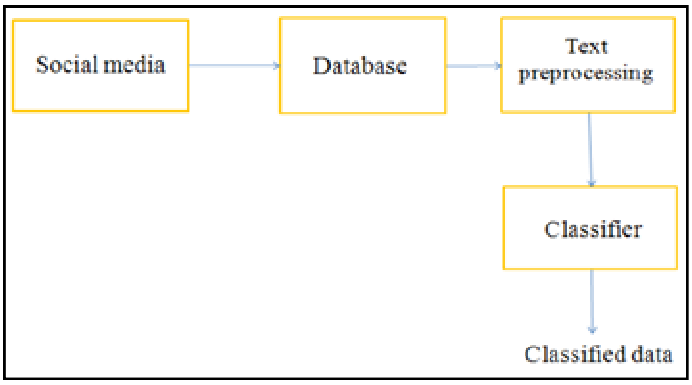

Fig. 2: System Architecture

A. Data Collection

It is challenging to collect social media data related to student's experiences because of the irregularity and diversity of the language used. Collected all the information from the comments that different students posted in social media. Also collected students email-id's to send the suggestions to their individual id's.

B. Text Preprocessing

- Negative words are useful for detecting negative emotion and issues. So we substituted words ending with "n't" and other common negative words (e.g., nothing, never, none, cannot) with "negtoken".
- Removed all words that contain non letter symbols and punctuations.
- 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, "soooocuuuteeee" is corrected to "So cute".
- Removed all stopwords such as an, a, the, between etc. Removing of stopwords improve efficiency.
- Performed stemming. Stemming is the process of reducing a word to a root, or simpler form.

C. Text Classification

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data) [1] [4].

In supervised machine learning techniques, labeled training documents are used. Naive Bayes, support vector machines and maximum entropy classifiers are the techniques used in this approach. In unsupervised machine learning techniques, there are no trained data. Based on the number of classes involved in the classification algorithms, there are binary classification and multi-class classification approaches. In binary classification, there are only two classes, while multi-class classification involves more than two classes. Both binary classification and multi-class classification are single-label classification systems. Single-label classification means each data point can only fall into one class where all classes are mutually exclusive. Multi-label classification, however, allows each data point to fall into several classes at the same time [5]. In the proposed work comments are classified using

Naive Bayes Multi-label classifier. This is a supervised machine learning technique.

Naive Bayes is a simple probabilistic classifier based on bayes theorem with strong independence assumptions between the features. The following are the basic procedures of the Naive Bayes multi-label classifier.

Suppose there are a total number of N words in the training document collection,

W= $\{w_1, w_2, \dots, w_N\}$

and a total number of L categories,

C= $\{c_1, c_2, \dots, c_L\}$

If a word w_n appears in a category c for $m_{w_n c}$ times, and appear in categories other than c for $m_{w_n c'}$ times, then based on the maximum likelihood estimation, the probability of this word in a specific category c is,

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

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

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

Suppose there are M documents in the document collection. For a document d_i in the testing set, there are K words $W_{d_i}= \{W_{i1}$, W_{i2} , …. ,W_{iK} $\}$ and W_{d_i} is a subset of W. According to Bayes Theorem, the probability that d_i belongs to category c is,

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

If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(c_1) =P(c_2) = . . . = P(c_L) and the probability that d_i belongs to other category is,

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

Here we choose a probability threshold T. If p(c|d_i)is larger than the probability threshold T, then d_i belongs to category c, otherwise, d_i does not belongs to category c. Then repeat this procedure for each category. If for a certain document, there is no category with a positive probability larger than T, we assign the one category with the largest probability to this document.

D. Suggestion And Feedback

After classification, finally we send the suggestions against their problems to their individual email-id's to those who need solution to their problem so that we provide privacy to the students also get feedback from the students in which how helpful our suggestions to them.

IV. Experimental Evaluation

Collected all the information from the comments that different students posted in social media website. In our case each comment is considered as a single document.

A. Text Preprocessing

Making sense out of social media content is a difficult task because of the highly informal and incoherent nature of the messages. So we preprocessed the text before training the classifier.

B. Development of Categories

There were no pre-defined categories of the data, so needed to explore what students were saying in the social media. Some of the problems faced by students include curriculum problems, heavy study load, study difficulties, imbalanced life, future and carrier worries, lack of gender diversity, sleep problems, stress, lack of motivation, physical health problems, nerdy culture, identity crisis, and others. These initial categories were developed in to 5 main categories: heavy study load, lack of social engagement, negative emotion, sleep problems, and diversity issues. Each category reflects one issue or problem that engineering students encounter in their learning experiences.

Table 1: examples of some comments under each category.

CATEGORY	COMMENTS
Heavy Study Load	back hurts from sitting at the desk studying for so long… so much of homework..
Lack of Social Engagement	sacrifice time to do homework miss party with friends..
Negative Emotion	No fun I feel myself dying
Sleep Problems	lots of work, can't sleep
Diversity Issues	only 5 girls in a class of 45 boys.. too many foreign lectures, unable to digest the subject

C. Classification

A classifier's job is to assign a Concept to an Instance. In order to know what Concept should be assigned to a particular Instance, a classifier reads a Training Set-a set of Instances that already have Concepts assigned to them. Upon loading those Instances, the classifier trains itself. Thereby it learns how to map a Concept to an Instance based on the assignments in the Training Set. The learning is "supervised" because the classifier undergoes a process of training, based on known classifications, and through supervision it attempts to learn the information contained in the training dataset. On using Naive Bayes approach to classify test data we get the following result.

Commonly used measure to evaluate the performance of classification models includes accuracy, precision, recall and the harmonic average between precision and recall-the F1 score [1] [4]. Positive documents are the documents of the main class of interest and negative documents are all other documents. There are four additional terms we need to know that are the building blocks used in computing evaluation measures.

- True positives (TP): These refer to the positive documents that were correctly labeled by the classifier. Let TP be the number of true positives.
- True negatives (TN): These are the negative documents that were correctly labeled by the classifier. Let TN be the number of true negatives.
- False positives (FP): These are the negative documents that were incorrectly labeled as positive. Let FP be the number of false positives.
- False negatives (FN): These are the positive documents that were mislabeled as negative. Let FN be the number of false negatives.

Accuracy a = $\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$

Precision p = $\frac{t_p}{t_p+f_p}$

Recall r = $\frac{t_p}{t_p+f_n}$

F1 Score = $\frac{2*p*r}{p+r}$

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. In the pattern recognition literature, this is also referred to as the overall recognition rate of the classifier, that is, it reflects how well the classifier recognizes documents of the various classes. Recall is also referred to as sensitivity or true positive rate, i.e., the proportion of positive tuples that are correctly identified. The precision and recall measures are widely used in classification. Precision can be thought of as a measure of exactness (i.e., what percentage of documents labeled as positive are actually such), whereas recall is a measure of completeness (what percentage of positive documents are labeled as such). The F measure is the harmonic mean of precision and recall. It gives equal weight to precision and recall.

The documents are classified using Naive Bayes multi-label classifier. Since it is a multi-label classifier, each document falls in to multiple categories at the same time. In this study, it is found that most students are largely struggling with negative emotions and not able to manage it successfully. Table 1 shows the values of true positive, true negative, false positive and false negative for varying number of documents. And table 2 shows the values of accuracy, precision, recall and f1 score for varying number of documents calculated using values of true positive, true negative, false positive and false negative in table 3.

Table 2: Parameters for calculating accuracy, precision, recall and F1 Score

No. of documents	True positive	True negative	False positive	False negative
30	7	23	5	0
50	17	33	4	2
100	23	77	7	5

Table 3: Evaluation Measures with Naive Bayes Classifier

No. of documents	Accuracy	Precision	Recall	F1 Score
30	0.85	0.5	1	0.66
50	0.8	0.85	0.8	0.82
100	0.86	0.76	0.82	0.78

V. Conclusion and Future Work

Most existing studies are either binary classification or multi class classification. The proposed system implemented a naive bayes multi-label classifier, which allowed one comment to fall into multiple categories at the same time. This study 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 large scale computational analysis of user-generated textual content. This study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering student's college experiences. Future work can consider positive aspects in student's learning experience. In the future work we can include more categories. Other possible future work could analyze student's generated content other than texts (e.g., images and videos). Use NLP techniques to get in-depth meaning of student posted contents.

VI. Acknowledgment

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

References

[1] *Jiawei Han, Micheline Kamber, Jian Pei, " Data Mining: Concepts and Techniques ", 3rd ed, Elsevier, 2012.*

[2] *R. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.*

[3] *R. Ferguson, "The State of Learning Analytics in 2012:A Review and Future Challenges," Technical Report KMI-2012-01, Knowledge Media Inst. 2012.*

[4] *Daniel T. Larose, "Data Mining: Methods and Models", Sharda Offset press, 2012.*

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

[6] *B. Pang, L. Lee, and S. Vaithyanathan, " Thumbs Up?: Sentiment Classification Using Machine Learning Techniques", Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing, vol. 10,pp. 79-86, 2002.*

[7] *H. Loshbaugh, T. Hoeglund, R. Streveler, and K. Breaux, "Engineering School, Life Balance, and the Student Experience", Proc. ASEE Ann. Conf. and Exposition, 2006.*