Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 5, Issue. 2, February 2016, pg.213 – 217

Social Media Mining To Analyse Students' Learning Experience

Ms. S. Aswini, Dr. Ilango Krishnamoorthy

Department of Computer Science and Engineering
Sri Krishna Collage of Engineering and Technology, Coimbatore, Tamil Nadu, India
14mg005@skcet.ac.in, ik@skcet.ac.in

Abstract— ***Students' casual posts and comments on social media (social network sites such as twitter) focuses into their educational experiences such as their problems, issues, suggestions about the students' study process. Some valuable information about student learning experiences can be inferred from the data gathered from such environments. It is very hard to analyze that information. Since the social media data keeps increasing in size it demands automation in data analysis. Then also the information inferred from those data needs human interpretation because it is the reflection of students' crisis. A work flow that combines qualitative analysis and large-scale data mining techniques is generated. We focused on students' posts to understand issues and problems in their educational experiences. Heavy work load, lack of awareness social activities, and sleeplessness are some problems that students face as they go through academic process. Based on these results, we started to implement a multi-label classification algorithm to classify posts reflecting students' problems.***

Keywords— social mining, text classification, naïve Bayes classifier.

I. Introduction

Social media sites such as twitter, Facebook provide great venues for students to share their experiences, vent emotion and stress, and seek social support. On various social media sites, students discuss and share their every day encounters in an informal and casual manner. The abundance of social media data provides opportunities to understand students' experiences, but also raises methodological difficulties in making sense of social media data for educational purposes. Just imagine the sheer data volumes, the diversity of internet slang, the unpredictability of location and timing of students posting on the web, as well as the complexities of students' experiences. Pure manual analysis cannot deal with the ever-growing scale of data, while pure automatic algorithms usually cannot capture in-depth meaning within the data.

Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, and classroom activities to collect data related to students' learning experiences. These methods are usually very time consuming, thus cannot be duplicated or repeated with high frequency. The scale of such studies is also usually limited. In addition, when prompted about their experiences, students need to reflect on what they were thinking and doing sometime in the past, which may have become obscured overtime.

II. NAÏVE BAYES

Naive Bayes assumes a particular probabilistic generative model for text. The model is a specialization of the mixture model, and thus also makes the two assumptions discussed there. Additionally, naive Bayes makes word independence assumptions that allow the generative model to be characterized with a greatly reduced number of parameters. The rest of this subsection describes the generative model more formally, giving a precise specification of the model parameters, and deriving the probability that a particular document is generated given its class label.

First let us introduce some notation to describe text. A document, d_i , is considered to be an ordered list of word events, $w_{di,1}$, $w_{di,2}$, ... We write $w_{di,k}$ for the word wt in position k of document di , where wt is a word in the vocabulary V = w1 , w2 , . . . , $w_{|V|}$.

When a document is to be generated by a particular mixture component, c_j , a document length, |di |, is chosen independently of the component. (Note that this assumes that document length is independent of class.3) Then, the selected mixture component generates a word sequence of the specified length. We furthermore assume it generates each word independently of the length.

Thus, we express the probability of a document given a mixture component in terms of its constituent features: the document length and the words in the document. Note that, in this general setting, the probability of a word event must be conditioned on all the words that precede it.

$$P\ (\ d_i|c_j;\theta) = P((w_{d_{i,1},\ldots,}\ w_{d_{i,|di|}})|c_j;\theta)= P(|d_i|) \prod_{k=1}^{|\,d_i\,|} P(w_{d_{i,k}}|C_j;\,\theta\,;\,w_{d_{i,q}},q<k)$$

Next we make the standard naive Bayes assumption: that the words of a document are generated independently of context, that is, independently of the other words in the same document given the class label. We further assume that the probability of a word is independent of its position within the document; thus, for example, the probability of seeing the word “homework” in the first position of a document is the same as seeing it in any other position. We can express these assumptions as:

$$P(w_{d_{i,k}}|C_j;\,\theta\,;\,w_{d_{i,q}},q<k) = P(w_{d_{i,k}}|C_j;\,\theta\,)$$

Combining these last two equations gives the naive Bayes expression for the probability of a document given its class:

$$P(w_{d_{i,k}}|C_j;\,\theta\,) = P(|d_i|) \prod_{k=1}^{|\,d_i\,|} P(w_{d_{i,k}}|C_j;\,\theta\,)$$

Thus the parameters of an individual mixture component are a multinomial distribution over words, i.e. the collection of word probabilities, each written θ wt $|c_j$, such that θ wt $|c_j = P(wt \mid c_j ; \theta)$, where t = {1, . . . , |V |} and t P(wt | c_j ; θ) = 1. Since we assume that for all classes, document length is identically distributed, it does not need to be parameterized for classification. The only other parameters of the model are the mixture weights (class prior probabilities), written θc_j, which indicate the probabilities of selecting the different mixture components. Thus the complete collection of model parameters, θ, is a set of multinomial and prior probabilities over those multinomial: $\theta = \{\theta wt\ |c_j : wt \in V,\ c_j \in C;\ \theta c_j : c_j \in C\}$.

III. RELATED WORKS

The authors, Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell, says that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. This is important because in many text classification problems obtaining training labels is expensive, while large quantities of unlabeled documents are readily available. He introduces an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation-Maximization (EM) and a naive Bayes classifier. The algorithm first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents. It then trains a new classifier using the labels for all the documents, and iterates to convergence. This basic EM procedure works well when the data conform to the generative assumptions of the model. However these assumptions are often violated in practice and poor performance can result. These authors present two extensions to the algorithm that improve Classification accuracy under these conditions: (1) a weighting factor to modulate the contribution of the unlabeled data, and (2) the use of multiple mixture components per class. Experimental results, obtained using text from three different real-world tasks, show that the use of unlabeled data reduces Classification error by up to 30%.

As the authors, Bo Pang and Lillian Lee examined the relation between subjectivity detection and polarity classification, showing that subjectivity detection can compress reviews into much shorter extracts that still retain polarity information at a level comparable to that of the full review. In fact, for the Naive Bayes polarity classifier, the subjectivity extracts are shown to be more effective in-put than the originating document, which suggests that they are not only shorter, but also "cleaner" representations of the intended polarity. They have shown that employing the minimum-cut framework results in the development of efficient algorithms for sentiment analysis. Utilizing contextual information via this frame-work can lead to statistically significant improvement in polarity-classification accuracy. Now these authors conclude stating the naïve Bayes is comparatively best.

Tina R. Patil, Mrs. S. S. Sherekar sets out to make comparative evaluation of classifiers naïve bayes and J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool. The experiments results shown in this paper are about classification accuracy, sensitivity and specificity. The results in the paper on this dataset also show that the efficiency and accuracy of j48 is better than that of Naïve bayes. But our problem has the multi level classifier. The system has numerous data values. So we use naïve Bayes to handle the everyday growing dataset with the multiple attribute values.

IV. IMPLEMENTATION

Traditionally there is no work flow to sense the data in the social media for the educational purpose and to integrate both qualitative analysis and large-scale data mining techniques. The system that has been started to develop is to monitor the student activity by allowing the necessary privacy. The issues and problems of the students can be found out through their informal conversation in the social media. If the issues are negative and serious then the alert has to be sent to the concerned staff and their parents. The issues, problems and the feedback of the students can be considered in the higher level decision making.

The system has seven steps to get the confined result. They are,

1. Data Collection
2. Data Sampling
3. Qualitative Data Analysis
4. Qualitative Result
5. Model Training And Evaluation
6. Model Adaption
7. Large Scale Data Analysis Result

In the fig. 4.1. Data is collected as a data set (i.e.,) the post collected from the community forum is collected as a sample and analyzed. The issues and problems of the students can be found out through their post. If the issues are negative and serious then the alert has to be sent to the concerned staff and their parents. The issues, problems and the feedback of the students can be considered in the higher level decision making.

The proposed system has following modules,

Computational Attribute Management

The attribute generation module fully operates on the Administrator end and it gathers the attribute data from admin for maintaining the dataset for future purpose. Once the attributes are generated it will be easy for analyze the thoughts of the student based on their given feedback from their portal. For all the attribute management is a dataset management process which holds the information regarding the keywords and the respective attributes for mining the student performance and learning experiences.

Qualitative Attribute Analysis

Once the attributes are generated the student effective mining portal is in other hand to collect the information from the student end. From in this port only the administrator can analyze the thoughts of the students. At each and every time the student twit with some data that all will be compared with the created attribute set and the mining process takes place for analyzing the qualitative summary of the student. In particular the qualitative analysis module helps the administrator to analyze the qualitative summary of the student performance.

Public Web Conversation Port

The theoretical foundation for the value of informal data on the web can be drawn from Goffman's theory of social performance. Although developed to explain face-to-face interactions, Goffman's theory of social performance is widely used to explain mediated interactions on the web today. 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 front stage, the relaxing atmosphere of back-stage usually encourages more spontaneous actions. Whether a social setting is front-stage or back-stage is a relative matter. For students, compared with formal classroom settings, social media is a 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. These conversations act as a zeitgeist for students' experiences.

Mining Student Conversation

From the diverse fields of many existing works researchers have analyzed the mining content to generate specific knowledge for their respective subject domains. For example, Gaffney analyzes tweets with hash tag Iran Election using histograms, user networks, and frequencies of top keywords to quantify online activism. Similar studies have been conducted in other fields including healthcare, marketing and athletics, just to name a few. Analysis methods used in these studies usually include qualitative content analysis, linguistic analysis, network analysis, and some simplistic methods such as word clouds and histograms. In this module, a classification model is built based on inductive content analysis. This model was then applied and validated on a brand new dataset. Therefore, not only the insights gained from one dataset are emphasized, but also the application of the classification algorithm to other datasets for detecting student problems. The human effort is thus augmented with large-scale data analysis.

Text Preprocessing

Many social mining users use some special symbols to convey certain meaning. For example, # is used to indicate a hashtag, @ is used to indicate a user account, and RT is used to indicate a re-tweet. Social Mining users sometimes repeat letters in words so that to emphasize the words, for example, "huuungryyy", "sooo muuchh", and "Monnndayyy". Besides, common stopwords such as "a, an, and, of, he, she, it", nonletter symbols, and punctuation also bring noise to the text. So the pre-processed texts are required completely for analysing the student mining data and through this module easily it will be attained.

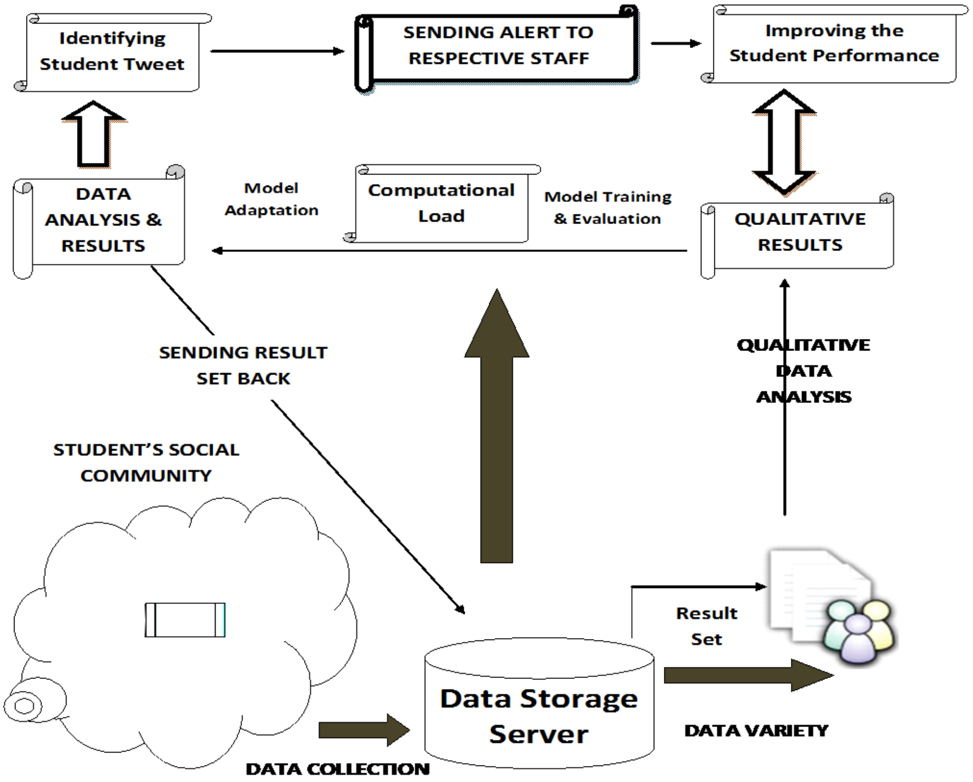
Identifying Student Tweet
SENDING ALERT TO RESPECTIVE STAFF
Improving the Student Performance
DATA ANALYSIS & RESULTS
Model Adaptation
Computational Load
Model Training & Evaluation
QUALITATIVE RESULTS
SENDING RESULT SET BACK
QUALITATIVE DATA ANALYSIS
STUDENT'S SOCIAL COMMUNITY
Result Set
Data Storage Server
DATA VARIETY
DATA COLLECTION

Fig 1: System Architecture

V. CONCLUSION

This paper 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. And also it can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' college experiences. In this project, the community like structure is planned between students and staff. Manually the admin has to collect the data from the students and post the alert message to the staff.

VI. FUTURE WORK

In future work, the automation of the finding the attribute in the post of the students, alerting the staffs will be the major part. The automation in correcting the emphasize words also take a major role. Analyzing the collective posts can also add as a significant part.

REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu and Krishna Madhavan, " Mining social media data for understanding students learning experience" , IEEE transaction on Learning Technologies, vol.7, no.3, Pp,16-22 July - September 2014

[2] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell, “Text Classification From Labeled And Unlabeled Documents Using EM”, Machine Learning, 39, Kluwer Academic Publishers. Printed In The Netherlands, Pp. 103–134, 2000.

[3] Bo Pang And Lillian Lee,”A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts”, Morgan & Clay Pool Publishers, Pp. 54-58, 2008.

[4] Tina R. Patil, Mrs. S. S. Sherekar,”Performance Analysis Of Naive Bayes And J48 Classification Algorithm For Data Classification”, J.sci.Education, Vol.86, No.1, Pp.7-15, 2000