Survey on Social Media using Twitter Data for Students Experience using Naïve Bayes Technique

Student, ME, Computer Engineering Department, SOCET

Harshit P Vora
M.E (C.E),
SOCET,
Ahmedabad,
Gujarat, India
.

Prof. Hiren Patel
Assistant Professor,
Computer/IT Department,
SOCET,
Ahmedabad,
Gujarat, India.

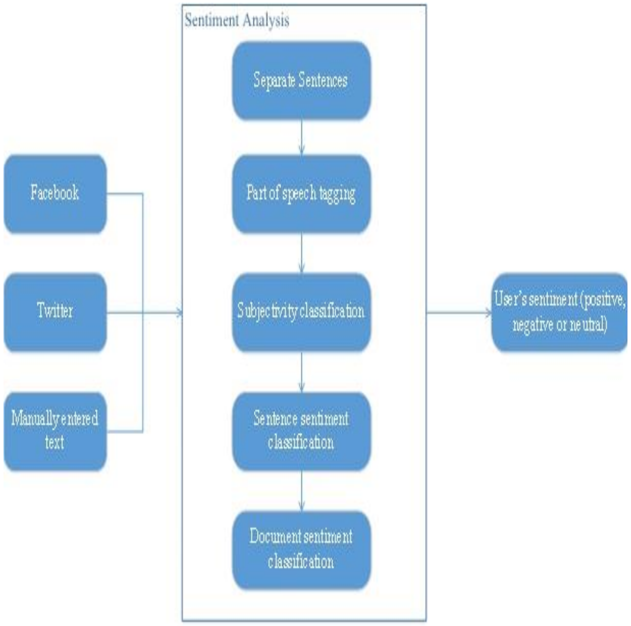
ABSTRACT

Students informal conversations on social media (e.g. Twitter, Face book) shed light into their educational experiences opinions, feelings, and concerns about the learning process. Data from such un-instrumented environments can provide valuable knowledge to inform student learning. Analysing such data, however, can be challenging. In this paper, we developed a workflow to integrate both qualitative analysis and large-scale data mining techniques [1]. We focused on engineering students' Twitter posts to understand issues and problems in their educational experiences. We first conducted a qualitative analysis on samples taken from about 25,000 tweets related to engineering students' college life. We found engineering students encounter problems such as heavy study load, lack of social engagement, and sleep deprivation. Based on these results, we implemented a multi-label classification algorithm to classify tweets reflecting students' problems. We then used the algorithm to train a detector of student problems from about 35,000tweets streamed at the geo-location of Purdue University [2]. This work, for the first time, presents a methodology and results that show how informal social media data can provide insights into students' experiences [6].

Keywords—Education, computers and education, social networking, web text analysis, twitter, multi label classifiers, Naïve Bayes.

INTRODUCTION

Social media sites such as Twitter, Face book, 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 andcasual manner. Students' digital footprints provide vast amount of implicit knowledge and a whole new perspective for educational researchers and practitioners to understand students' experiences. Students' feedback can highlight different issues students may have with a lecture. An example of this is the students not understanding a specific example. Analysing feedback in real-time, howevertime consuming is and stressful [4]. We propose to address this problem by creating a system that analyses students' feedback in real-time and then presents the results to the lecturer. To create such as system, sentiment analysis can be used.


Sentiment Analysis
Separate Sentences
Part of speech tagging
Subjectivity classification
Sentence sentiment classification
Document sentiment classification
Facebook
Twitter
Manually entered text
User's sentiment (positive, negative or neutral)

[I]Sentiment Analysis Process

Sentiment analysis is an application of natural language processing, computational linguistics and text analytics that identifies and retrieves sentiment polarity from the text by studying the opinion. Sentiment polarity is usually either positive or negative, although sometimes neutral is included. Previous research has shown that sentiment analysis is more effective when applied to specific domains [7]. Sentiment analysis in the educational domain has mainly been focused on e-learning, with little research done on classroom feedback [8]. Although e-learning and classroom education may seem similar, they differ in the types of interactions between the students and the lecturers and in the fact that the lecturer should respond to the students' feedback in real-time. Feedback from students in the classroom settings is different from distant learners due to the different situations and issues students may have. E-learning students can have issues such as lack of interaction. Engineering schools and departments have long been struggling with student recruitment and retention issues [8]. Engineering graduates constitute a significant part of the nation's future workforce and have a direct impact on the nation's economic growth and global competency [9].

To the best of our knowledge, sentiment analysis has not been applied for analysing students' classroom feedback before. Consequently, there is need of investigating different models and look at the best combination of pre-processing methods, features and machine learning techniques to create the best-suited model for our purpose. In this paper, we investigate the following aspects:

I. LITERATURE REVIEW

In this paper, they have developed a workflow to integrate both qualitative analysis and large scale data mining techniques. They have focused on engineering students' twitter posts to understand issues and problems in their educational experiences. They have first conducted a qualitative analysis on samples taken from about 25,000 tweets related to engineering students' college life[1].

In this paper, they found that engineering students encounter problems such as heavy study loads, lack of social engagement, and sleep deprivation. Based on this result, they implemented a multi-label classification algorithm to classify tweets reflecting students' problems. This work for the first time presents a methodology and results that show how informal social media data can provide insights into students' experience[2].In this paper, they found that engineering students encounter problems such as heavy study loads, lack of social engagement, and sleep deprivation. Based on this result, they implemented a multi-label classification algorithm to classify tweets reflecting students' problems. This work for the first time presents a methodology and results that show how informal social media data can provide insights into students' experience.Social networks are around as long as humans organized themselves into groups [4]. The digital age introducing social media gave social networks an unprecedented impulse to flourish. Meanwhile, quite a few studies have been carried out on social media and social networks. In order to get a comprehensive view of the literature, we limit our focus on the most widely used online social network for scientific analysis, Twitter. Most of the research in this field is based on data gathered from Twitter. Although research has been done based on other social networks and media like Face book and YouTube, this work has not led to new insights, new approaches or novel contributions, as compared to Twitter [8].

II) TECHNIQUESUSED FOR TWITTER DATA.

Classification is the separation or ordering of objects into classes. There are two phases in classification algorithm: first, the algorithm tries to find a model for the class attribute as a function of other variables of the datasets. Next, it applies previously designed model on the new and unseen datasets for determining the related class of each record Text classification is to automatically assign the texts into the predefined categories. Text categorization mostly depends on the information retrieval technique such as indexing, inductive construction of classifiers and evaluation technique. In this machine learning, classifier learns how to classify the categories of documents based on the features extracted from the set of training data. Social content mining can be done on unstructured data such as text. Mining of unstructured data have hidden information and Text Mining is extraction of previously unknown information extracting information from different text sources. Social content mining requires application of data mining and text mining techniques [3].

1) Bayesian Classifier

The most commonly used classifier for Text classification. Basic idea behind this classifier is to find probability that to which class this document belong. Using this, we can understand the profiles by the feedback collected from various Social media sites. It is simple, but often outperforms more sophisticated classification methods. Maximum Likelihood estimates the parameters for the models. It requires small number of training to estimate the parameters. It Works well and efficiently in supervised learning. Here, the rank order of the pages will be rated. Text classification is based on calculating the posterior probability of the documents present in the different classes. Naïve Bayes is based on Bayesian theorem with independence feature selection [5]. Naïve Bayesian classification is used for anti- spam filtering technique. It is divided in two different phases. The first phase has been functional for training set of data and the second phase employs the classification phase [9].

2) K-nearest neighbor

K-NN classifier works on principle that is the points (documents) that are close in the space belong to the same class. It calculates similarity between test document and each neighbor. It is a case-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measures. Many applications use this method because of its effectiveness, non-parametric and easy to implementation properties. However the classification time is long and difficult to find optimal value of k. The best choice of k depends upon the data. A good k can be selected by various heuristic techniques[1].

3) Support vector machine

SVM (Support Vector Machines) is one of the most used and accurate classifiers in many machine learning tasks, but our comparison experiment shows that Naïve Bayes exceedsSVM in this study. We first implemented a linear multi-label SVM using the LibSVM library with the one-versus-all Heuristic.We applied weight of loss parameters that are proportional to the inverse of the percentages of tweets in or not in each category to account for the imbalanced categories. However, with thesame training and testing data sets as in the above section, this one-versus-all SVM multi- label classifier classified all tweets into not in the categoryfor all categories. So we got empty label sets for all tweets.

Then we applied the same training and testing data sets as above to an advanced SVM variation named Max- Margin Multi-Label (M3L) classifier. M3L is a state-of-the art multi-label classifier. Different from the oneversus- all heuristic, which assumes label independence, this classifier takes label correlation into consideration[2].

We used the executable file of this algorithm provided by the author. The performance is better than the simplistic one-versus-all SVM classifier, but still not as good as the Naive Bayes classifier. Table 4 and Fig. 3 show the evaluation measures using (Max Margin Multi-Label (M3L) classifier[4].

4) Decision Tree

According to the datasets by information extraction, a decision tree constantly updated data to update the decision tree, and then generate the understandable rules. The experiment proves that it is feasible to realize the Web information extraction based on the decision tree [1].

Decision Tree Technology, which has three courses.

a) Construct wrapper.

b) Decision tree building process, a rough decision tree will be constructed based on an algorithm.

c) Decision tree refinement process and to automatically extracted knowledge or rules.

Classifier	Advantage	Disadvantage
Naïve Bayes	Easy Implement. Easy Computation	Perform Poor.
Support Vector Machine(SVM)	Capture the Data Inherent better.	Parameter tuning.
k-nearest	Effective. Non-Parametric.	Classification time is too long.
Decision Tree	Easy To understand. Easy to generate rule	Does not handle continue variable well.

II. ISSUE IN PREVIOUS APPROACH

In this section, Naïve Bayes multi-label classifier is used to detect engineering student problems from the Purdue dataset. There were 35,598 unique tweets in the Purdue tweet collection. We took a random sample of 1,000 tweets, and found no more than 5% of these tweets were discussing engineering problems. Our purpose here was to detect the small number of tweets that reflects engineering students' problems. The differences between #engineering Problems dataset and Purdue dataset is that the latter contains much smaller number of positive samples to be detected, and its "others" category has more diverse content[5].

III. SYSTEM OVERVIEW

First, not all students are active on Twitter, so we may only find the ones who are more active and more likely to expose their thoughts and feelings.

Second, the fact that the most relevant data we found on engineering students' learning experiences involve complaints, issues, and problems does not mean there is no positive aspects in students' learning experiences. We Did find a small number of tweets that discuss the good things about being engineering students such as those using hash tag #engineering Perks. We chose to focus on the problems in this paper because these could be the most informative for improvement of education quality.

Other possible future work could analyse students' generated content other than texts (e.g. images and videos), on social media sites other than Twitter (e.g. face book, Tumbler and YouTube). Future work can also extend to students in other majors and other institutions.

III. CONCLUSIONS

Our study is beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analysing social data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content.
Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' college experiences[1].

As an initial attempt to instrument the uncontrolled social media space, we propose many possible directions for future work for researchers who are interested in this area hope to see a proliferation of work in this area in the near future. We advocate that great attention needs to be paid to protect students' privacy when trying to provide good education and services to them[9].

I) REFERENCES

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

[2] 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.

[3] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens. Streveler, and K. Smith, "Academic pathways study: Processes and realities," in *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*, 2008.

[4]S. Cetin as, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez, "Micro blogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Micro bloggingSupported Classroom," *Learning Technologies, IEEE Transaction son*, vol. 4, no. 4, pp. 292–300, 2011.

[5] R. Baker and K. Yacef, "The state of educational data mining in2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[6] K. Nirmala, S. Satheesh kumar and Dr. J. Vellingiri "A Survey on Text categorization in Online Social Networks" International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 9, September 2013.

[7] J. Yang and S. Counts, "Predicting the speed, scale, and rangeof information diffusion in twitter," *Proc. ICWSM*, 2010.

[8] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 23–32.

[9] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," presented at *The*

International AAAI Conference on Weblogs and Social Media (ICWSM),2012.