Mining Social Media for Understanding Students' Learning Experiences

L.R.Jeevitha [1] R.Priyanka[2] P.Deepa[3] A.N.Sasikumar [4]

PG Student[1,2] Associate Professor[3,4]

[1, 2,3,4] Department of M.C.A., Panimalar Engineering College, Chennai, Tamilnadu, India

jeee1993@gmail.com[1] , mca_deepa@yahoo.com[3]

Abstract- Students' informal conversations on social media (e.g. Twitter, Facebook) 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. Analyzing such data, however, can be challenging. The complexity 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 developed a workflow to integrate both qualitative analysis and large-scale data mining techniques. 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 35,000 tweets streamed at the geo-location of Purdue University. This work, for the first time, presents a methodology and results that show how informal social media data can provide insights into students' experiences.

Key Terms-Education, computers and education, social networking, web text analysis

1 INTRODUCTION

Automated prediction of trends and behaviors: Mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing. Our purpose is to achieve deeper and finer understanding of students' experiences especially their learning-related issues and problems. To determine what student problems a tweet indicates is a more complicated task than to determine the sentiment of a tweet even for a human judge. Therefore, our study requires a qualitative analysis, and is impossible to do in a fully unsupervised way. Sentiment analysis is, therefore, not applicable to our study. In our study, we implemented a multi-label classification model where we allowed one tweet to fall into multiple categories at the same time. Our work extends the scope of data-driven approaches in education such as learning analytics and educational data mining. Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, 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 emerging field of learning analytics and educational data mining has focused on analyzing structured data obtained from course management systems (CMS), classroom technology usage, or controlled online learning environments to inform educational decision making. However, to the best of our knowledge, there is no research found to directly mine and analyze student- posted content from uncontrolled spaces on the social web with the clear goal of understanding students' learning experiences. The drawbacks are In our study, through a qualitative content analysis, we found that engineering students are largely struggling with the heavy study load, and are not able to manage it successfully. Heavy study load leads to many consequences including lack of social engagement, sleep problems, and other psychological and physical health problems. Our work is only the first step towards revealing actionable insights from student-generated content on social media in order to improve education quality. We extend the proposed algorithm which analysis the student's learning experiences by giving solutions to their problems. The suggested solution is forwarded to the student's individual email-ids to attain the privacy of student and for improving security a novel secure algorithm called BIRCH is proposed. Finally we get the feedback from the students about solution provided and comparison graph is generated.

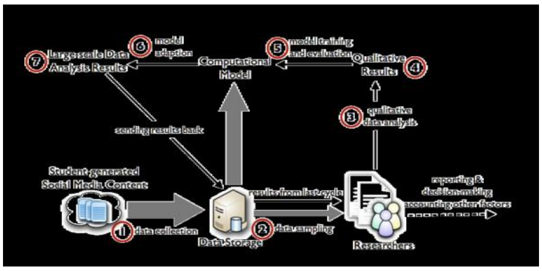
2 RELATED WORKS

The research goals of this study are
1) To demonstrate a workflow of social media data sense-making for educational purposes, integrating both qualitative analysis and large-scale data mining techniques and
2) To explore engineering students' informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences. 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. Amongst these topics, tweet classification is most relevant to this study. Popular classification algorithms include Naïve Bayes, Decision Tree, Logistic Regression, Maximum Entropy, Boosting, and Support Vector Machines (SVM). Most existing studies found on tweet classification are either binary classification on relevant and irrelevant content, or multi-class classification on generic classes such as news, events, opinions, deals, and private messages. 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). We chose to focus on engineering students' posts on Twitter about problems in their educational experiences mainly because:
1. Engineering schools and departments have long been struggling with student recruitment and retention issues. 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.
2. Based on understanding of issues and problems in students' life, policymakers and educators can make more informed decisions on proper interventions and services that can help students overcome barriers in learning.
3. Twitter is a popular social media site. Its content is mostly public and very concise (no more than 140 characters per tweet). Twitter provides free APIs that can be used to stream data. Therefore, we chose to start from analyzing students' posts on Twitter.

3 PROPOSED WORK

The research goals of this study are
1) To demonstrate a workflow of social media data sense-making for educational purposes, integrating both qualitative analysis and large-scale data mining techniques as illustrated in Fig. 1; and
2) To explore engineering students' informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences.
We chose to focus on engineering students' posts on Twitter about problems in their educational experiences mainly because:
1. Engineering schools and departments have long been struggling with student recruitment and retention issues. 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
2. Based on understanding of issues and problems in students' life, policymakers and educators can make more informed decisions on proper interventions and services that can help students overcome barriers in learning.
3. Twitter is a popular social media site. Its content is mostly public and very concise (no more than 140 characters per tweet). Twitter provides free APIs that can be used to stream data. Therefore, we chose to start from analyzing students' posts on Twitter.

3.1 System Architecture


7 Large-scale Data Analysis Results
6 model adaption
Computational Model
5 model training and evaluation
Qualitative Results 4
3 qualitative data analysis
sending results back
Student generated Social Media Content
1 data collection
Data Storage
results from last cycle
2 data sampling
Researchers
reporting & decision-making accounting other factors

4 MODULE DESCRIPTIONS

4.1 Login

In this module, the user is login to the social website. So that, we can see the posts by the engineering students.

4.2 Data collection

Collects all the information from the different students posted their comments in social website twitter.

In this we also collect students email-id's to send the suggestions to their individual id's.

4.3 Data clustering

In this module, the raw data is clustered by using clustering algorithm. This algorithm starts with single cluster. Every point in a database is a cluster. Then it groups closest points into separate clusters, and continues until only one cluster remains. The computation of clusters calculated with help of distance matrix. The algorithm generates cluster feature tree while scanning the dataset. Each entry in the CF tree represents the cluster of objects and is characterized by triple (N, LS, SS).

4.4 Data classification

After clustering the data in different clusters based on the content, we use Naïve Bayes classification algorithm. One popular way to implement multi-label classifier is to transform the multi-label classification problem into multiple single-label classification problems. One simple transformation method is called one-versus-all or binary relevance. The basic concept is to assume independence among categories, and train a binary classifier for each category. All kinds of binary classifier can be transformed to multi-label classifier using the one-versus-all heuristic.

4.5 Suggestions and feedback

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

5 IMPLEMENTATION

5.1 Data Collection

It is challenging to collect social media data related to students' experiences because of the irregularity and diversity of the language used. We searched data using an educational account on a commercial social media monitoring tool named Radian6.The Twitter APIs can also be configured to accomplish this task, which we later used to obtain the second dataset. The search process was exploratory. We started by searching based on different Boolean combinations of possible keywords such as *engineer, students, campus, class, homework, professor,* and *lab*. We then expanded and refined the keyword set and the combining Boolean logic iteratively. The Boolean search logic grew very complicated eventually, but the dataset still contained about 35% noise. Also, given that the dataset was so small, we seemed to have ruled out many other relevant tweets together with the spam and irrelevant tweets.

From the limited number of relevant tweets we retrieved, we found a Twitter hashtag #engineeringProblems occurring most frequently. Students used the hashtag #engineeringProblems to post about their experiences of being engineering majors. This was the most popular hashtag specific to engineering students' college life based on the data retrieved using the Boolean terms.

Using Radian6, we streamed tweets containing this hashtag for about 14 months (421 days) from November 1st, 2011 to December 25th, 2012. In total, we collected 25,284 tweets with the hashtag #engineeringProblems posted by from 10,239 unique Twitter accounts.

Counting re-tweets, replies, and mentions, a total of 12,434 unique user accounts were involved. After removing duplicates caused by re-tweeting, there were 19,799 unique tweets in this dataset. We also identified several other much less popular but relevant hashtags such as #ladyEngineer, #engineering- Majors, #switchingMajors, #collegeProblems, and #nerdstatus. As a side note for future work, these hashtags can also be used to retrieve data relevant to college students' experiences.

To demonstrate the application of the classification algorithm, we obtained another new dataset using the geocode of Purdue University West Lafayette (40.428317, -86.914535) with a radius of 1.3 miles to cover the entire campus. From February 5th to April 17th, 2013, we obtained 39,095 tweets using the Twitter search API. These tweets came from 5,592 unique user accounts.There were 35,598 unique tweets after removing duplicates. One reason we chose Purdue as an example is that it is a large public university with a strong engineering student base.

Over 27% (10,533/39,000) of the students at the West Lafayette campus are enrolled in the College of Engineering during the 2012-2013 academic year. Nevertheless, the general approach we used can be applied to any institution and students in any major.

5.2 Inductive Content Analysis

Because social media content like tweets contain a large amount of informal language, sarcasm, acronyms, and misspellings, meaning is often ambiguous and subject to human interpretation. Rost et. al 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. We concur with this argument, as we found no appropriate unsupervised algorithms could reveal in-depth meanings in our data. For example, LDA (Latent Dirichlet Allocation) is a popular topic modeling 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.

There were no pre-defined categories of the data, so we needed to explore what students were saying in the tweets. Thus, we first conducted an inductive content analysis on the #engineeringProblems dataset. Inductive content analysis is one popular qualitative research method for manually analyzing text content. Three researchers collaborated on the content analysis process.

5.3 Development of Categories

The lens we used in conducting the inductive content analysis was to identify what are the major worries, concerns, and issues that engineering students encounter in their study and life. Researcher A read a random sample of 2,000 tweets from the 19,799 unique #engineeringProblems 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, stress, lack of motivation, physical health problems, nerdy culture, identity crisis, and others. These were developed to identify as many issues as possible, without accounting for their relative significances. Researcher A wrote detailed descriptions and gave examples for each category and sent the codebook and the 2,000-tweet sample to researchers B and C for review. Then, the three researchers discussed and collapsed the initial categories into five prominent themes, because they were themes with relatively large number of tweets. We found that many tweets could belong to more than one category. For example, "*This could very well turn into an all nighter...f*** you lab report #nosleep*" 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. Too much stuff. Way too complicated. No fun*" falls into heavy study load, and negative emotion at the same time. So one tweet can be labeled with multiple categories. This is a multi-label classification as opposed 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.

5.4 Classification Results

From the inductive content 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 set were found in the training data set.

6 CONCLUSION

Our 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. Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' college experiences. 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, good education and services to them. In the future, which analysis the student's learning experiences by giving solutions to their problems. The suggested solution is forwarded to the student's individual email-ids to attain the privacy of student and for improving security by a novel secure algorithm. Finally get the feedback from the students about solution provided to generate comparison graph.

7 FUTURE WORK

This study explores the previously uninstrumented space on Twitter in order to understand engineering students' experiences, integrating both qualitative methods and large-scale data mining techniques. In our study, through a qualitative content analysis, we found that engineering students are largely struggling with the heavy study load, and are not able to manage it successfully. Heavy study load leads to many consequences including lack of social engagement, sleep problems, and other psychological and physical health problems. Many students feel engineering is boring and hard, 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, we implemented and evaluated a multi-label classifier to detect engineering student problems from Purdue University. This detector can be applied as a monitoring mechanism to identify at-risk students at a specific university in the long run without repeating the manual work frequently. Our work is only the first step towards revealing actionable insights from student-generated content on social media in order to improve education quality. There are a number of limitations, which also lead to many possible directions for future work.

REFERENCES

[1] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *Educause 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,
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.

[4] C. J. Atman, S. D. Sheppard, J. Turns, R. S. Adams, L. Fleming, R. Stevens, R. A. Streveler, K. Smith, R. Miller, L. Leifer, K. Yasuhara, and D. Lund, "Enabling engineering student success: The final report for the Center for the Advancement of Engineering Education," Morgan & Claypool Publishers, Center for the Advancement of Engineering Education, 2010.

[5] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," *Knowledge Media Institute, Technical Report KMI-2012-01*, 2012.