THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Mining Students Social Data to Understand Their Learning Experience

Janani Sankari.M
Assistant Professor,Department of Information Technology, Jeppiaar Engineering College,Chennai,India
Shoba L.K
Assistant Professor, Department of Information Technology, Jeppiaar Engineering College,Chennai,India

Abstract:
Students informal conversation on social network (Eg.Facebook, Twitter)can be used to know their educational experiences such as opinions, concerns and learning process.These data from such sites will help us to know about students learning and their difficulties.The complexity is that we need human interpretation in analysis of the data.In this paper, we developed a workflow to integrate both qualitative analysis andlarge-scale data mining techniques. We focused on engineering students' Twitter posts to understand issues and problems intheir educational experiences. We found engineering students encounter problems such as heavy study load, lack of socialengagement, and sleep deprivation. Based on these results, we implemented a multi-label classification algorithm to classifytweets reflecting students' problems. We then used the algorithm to train a detector of student problems from about 35,000tweets. This work, for the first time, presents a methodology and results thatshow how informal social media data can provide insights into students' experiences.

Keywords: *Education, computers and education, social networking, web text analysis*

1. Introduction

Social media sites such as Twitter, Facebook, andYouTube provide great venues for students to sharejoy and struggle, vent emotion and stress, and seek socialsupport. On various social media sites, students discussand share their everyday encounters in an informal and casual manner. Students' digital footprints provide vastamount of implicit knowledge and a whole new perspectivefor educational researchers and practitioners to understandstudents' experiences outside the controlledclassroom environment. This understanding can informinstitutional decision-making on interventions for at-riskstudents, improvement of education quality, and thusenhance student recruitment, retention, and success. The abundance of social media data provides opportunitiesto understand students' experiences, but also raisesmethodological difficulties in making sense of social mediadata for educational purposes. Just imagine the sheerdata volumes, the diversity of Internet slangs, the unpredictabilityof locations, and timing of students posting onthe web, as well as the complexity of students' experiences.Pure manual analysis cannot deal with the ever-growing scale of data, while pure automatic algorithmsusually cannot capture in-depth meaning within the data. Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, classroomactivities to collect data related to students' learning experiences. These methods are usually very timeconsuming, thus cannot be duplicated or repeated withhigh frequency. The scale of such studies is also usuallylimited. In addition, when prompted about their experiences, students need to reflect on what they were thinkingand doing sometime in the past, which may have becomeobscured over time. The emerging field of learning analytics and educationaldata mining has focused on analyzing structureddata obtained from course management systems (CMS),classroom technology usage, or controlled online learningenvironments to inform educational decision-making ,.However, to the best of our knowledge, thereis 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 research goals of this study are 1) to demonstratea workflow of social media data sense-making for educational purposes, integrating both qualitative analysis andlarge-scale data mining techniques as illustrated in Fig. 1; and 2) to explore engineering students' informal conversationson Twitter, in order to understand issues andproblems students encounter in their learning experiences.
We chose to focus on engineering students' posts onTwitter about problems in their educational experiences mainly because:

- Engineering schools and departments have longbeen struggling with student recruitment and retention issues. Engineering graduates constitute a significant part of the nation's future workforceand have a direct impact on the nation's economicgrowth and global competency . The workflow we developed for making sense of social media data integrates qualitative analysis and data mining algorithms. Thewidth of gray arrows represents data volumes – wider indicates more data volume. Black arrows represent data analysis, computation, andresults flow. The dashed arrows represent the parts that do not concern the central work of this paper. This workflow can be an iterative cycle.

- Based on understanding of issues and problems instudents' life, policymakers and educators canmake more informed decisions on proper interventionsand services that can help students overcome barriers in learning.
- Twitter is a popular social media site. Its content ismostly public and very concise (no more than 140characters per tweet). Twitter provides free APIsthat can be used to stream data. Therefore, we chose to start from analyzing students' posts onTwitter.In this paper, we went through an exploratory processto locate the relevant data and relevant Twitter hashtags(a Twitter hashtag is a word beginning with a # sign,used to emphasize or tag a topic). We collected 25,284tweets using the hashtag #engineeringProblems over a period of 14 months, and a second dataset of 39,095tweets using the geo-code (longitude and latitude) ofPurdue University, West Lafayette. This corresponds tostep 1 Three researchers conducted an inductive contentanalysis on samples of the #engineeringProblems dataset,which corresponds to steps 2 and 3 in Fig. 1.
- wefound that major problems engineering students encounter
  in their learning experiences fall into several prominentcategories. Based on these categories, we implementeda multi-label Naïve Bayes classification algorithm.We evaluated the performance of the classifier by comparingit with other state-of-the-art multi-label classifiers
- Second, the paper providesdeep insights into engineering students' educational experiencesas reflected in informal, uncontrolled environments.Many issues and problems such as study-life balance,lack of sleep, lack of social engagement, and lack ofdiversity clearly emerge. These could bring awareness toeducational pedagogy, policy-making, and educationalpractice.The remainder of this paper is organized as follows:the next section reviews theory of public discourse online,related work on text classification techniques used foranalyzing tweets, and data-driven approaches in education.Section 3 describes the data collection process (step 1in Fig. 1). Section 4 details the inductive content analysisprocedures and categories identified (steps 2, 3, and 4).Section 5 details the implementation of the Naïve Bayesmulti-label classifier and the evaluation results (step 5). Insection 6, we show the comparison results of Naïve Bayesclassifier with the popular classifier—Support Vector Machines(SVM) and one of its variations Max-Margin Multi-Label (M3L) classifier. This is an additional of evaluationof the classifier in step 5. In section 7 we apply the NaïveBayes classifier to the Purdue dataset in order to demonstrateits application in detecting students' problems at aspecific university (steps 6 and 7). Section 8 discusses thelimitations and possible future work, and section 9 concludes this study.

2. Related Work

2.1. Public Discourse on the Web

The theoretical foundation for the value of informal dataon the web can be drawn from Goffman's theory of social performance .Although developed to explain face-tofaceinteractions, 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 frontstage,the relaxing atmosphere of back-stage usually encouragesmore spontaneous actions. Whether a social settingis front-stage or back-stage is a relative matter. For students, compared with formal classroom settings, socialmedia 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. Inthis sense, the data collected from online conversation may be more authentic and unfiltered than responses toformal research prompts. These conversations act as azeitgeist for students' experiences.Many studies show that social media users may purposefully manage their online identity to "look better"than in real life .Other studies show that there isa lack of awareness about managing online identityamong college students ,and that young people usuallyregard social media as their personal space to hangout with peers outside the sight of parents and teachers. Students' online conversations reveal aspects of theirexperiences that are not easily seen in formal classroomsettings, thus are usually not documented in educationalliterature.The abundance of social media data provides opportunitiesbut also presents methodological difficulties foranalyzing large-scale informal textual data. The next sectionreviews popular methods used for analyzing Twitterdata.

2.2. Mining Twitter Data

Researchers from diverse fields have analyzed Twittercontent to generate specific knowledge for their respectivesubject domains. For example, Gaffney analyses tweets with hashtag #iranElection using histograms, usernetworks, and frequencies of top keywords to quantifyonline activism. Similar studies have been conducted inother fields including healthcare, marketing, athletics , just to name a few. Analysis methods used inthese studies usually include qualitative content analysis,linguistic analysis, network analysis, and some simplisticmethods such as word clouds and histograms. In ourstudy, we built a classification model based on inductive content analysis. This model was then applied and validatedon a brand new dataset. Therefore, we emphasize not only the insights gained from one dataset, but also theapplication of the classification algorithm to other datasets for detecting student problems. The human effort isthus augmented with large-scale data analysis.Below we briefly review studies on Twitter from thefields of data mining, machine learning, and natural languageprocessing. These studies usually have more emphasison statistical models and algorithms. They cover awide range of topics including information ropagation and diffusion, popularity prediction, event detection ,topic discovery and tweet classification , to name a few.Amongst these topics, tweet classification is most relevantto our study.Popular classification algorithms include Naïve Bayes,Decision Tree, Logistic Regression, Maximum Entropy,Boosting, and Support Vector Machines (SVM). Based onthe number of classes involved in the classification algorithms,there are binary classification and multi-class classificationapproaches. In binary classification, there areonly two classes, while multi-class classification involvemore than two classes. Both binary classification and multi-class classification are single-label classification systems.Single-label classification means each data point can onlyfall into one class where all classes are mutually exclusive.Multi-label classification, however, allows each data pointto fall into several classes at the same time.Most existing studies found on tweet classification areeither binary classification on relevant and irrelevant content, or multi-class classification on generic classessuch as news, events, opinions, deals, and private messages.Sentiment analysis is another very popularthree-class classification on positive, negative, or neutral emotions/opinions. Sentiment analysis is very usefulfor mining customer opinions on products or companiesthrough their reviews or online posts. It finds wide adoptionin marketing and customer relationship management(CRM). Many methods have been developed to mine sentimentfrom texts. For example, both Davidov et al. and Bhayani et al.use emoticons as indicators to providenoisy labels to the tweets thus to minimize humaneffort. However, in the case of this paper, only knowingthe sentiment of student-posted tweets does not providemuch actionable knowledge on relevant interventionsand services for students. Our purpose is to achievedeeper and finer understanding of students' experiencesespecially their learning-related issues and problems. Todetermine what student problems a tweet indicates is amore complicated task than to determine the sentiment ofa tweet even for a human judge. Therefore, our studyrequires 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 classificationmodel where we allowed one tweet to fall into multiple categories at the same time. Our classification wasalso at a finer granularity compared with other generic classifications. Our work extends the scope of data-drivenapproaches in education such as learning analytics andeducational data mining.

3. Data Collection

It is challenging to collect social media data related tostudents' experiences because of the irregularity and diversity of the language used. We searched data using aneducational account on a commercial social media monitoring tool named Radian6 The Twitter APIs can also be configured to accomplishthis task, which we later used to obtain the seconddataset. The search process was exploratory. We startedby searching based on different Boolean combinations ofpossible keywords such as engineer, students, campus, class,homework, professor, and lab. We then expanded and refinedthe keyword set and the combining Boolean logiciteratively. The Boolean search logic grew very complicatedeventually, but the dataset still contained about 35%noise (during the month of November 2011, we retrieved179 tweets, in which 63 were irrelevant to college students).Also, given that the dataset was so small, weseemed to have ruled out many other relevant tweets togetherwith the spam and irrelevant tweets

3.1. Development of Categories

The lens we used in conducting the inductive contentanalysis was to identify what are the major worries, concerns, and issues that engineering students encounter intheir study and life. Researcher A read a random sampleof 2,000 tweets from the 19,799 unique #engineeringProblemstweets, and developed 13 initial categories including: curriculum problems, heavy study load, study difficulties,imbalanced life, future and carrier worries, lack ofgender diversity, sleep problems, stress, lack of motivation,physical health problems, nerdy culture, identitycrisis, and others. These were developed to identify asmany issues as possible, without accounting for their relativesignificances. Researcher A wrote detailed descriptionsand gave examples for each category and sent the codebook and the 2,000-tweet sample to researchers Band C for review. Then, the three researchers discussedand collapsed the initial categories into five prominent themes, because they were themes with relatively largenumber of tweets.

3.2. Heavy Study Load

Our analyses show that, classes, homework, exams, andlabs dominate the students' life. Libraries, labs, and theengineering building are their most frequently visitedplaces. Some illustrative tweets are "Study over 30 hoursfor a test", "so much homework, so little time", and "C++CAE project due Tuesday, Mfg project Wednesday, 25 PageTech Repot Wednesday + heavy homework load. Huzzah", and"homework never stops".Students express a very stressful experience in engineering.Not being able to manage the heavy study loadleads to consequences such as lack of social engagement,lack of sleep, stress, depression, and some health problems.

3.3. Sleep Problems

Our analyses find that sleep problems are widely commonamong engineering students. Students frequentlysuffer from lack of sleep and nightmares due to heavy study load and stress. For example, "Napping in the commonroom because I know I won't sleep for the next three days","If I don't schedule in sleep time, it doesn't happen", and "Iwake up from a nightmare where I didn't finish my physics labon time". Chronic lack of sleep or low-quality sleep canresult in many psychological and physical health problems,therefore this issue needs to be addressed.

4. Naïve Bayes Multi-Label Classifier

We built a multi-label classifier to classify tweets based onthe categories developed in the previous content analysisstage. There are several popular classifiers widely used indata mining and machine learning domain. We found NaïveBayes classifier to be very effective on our dataset comparedwith other state-of-the-art multi-label classifiers.

4.1. Text Pre-processing

Twitter 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. Twitter 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", nonlettersymbols, and punctuation also bring noise to thetext. So we pre-processed the texts before training the
classifier:

- We removed all the #engineeringProblemshashtags. For other co-occurring hashtags, we onlyremoved the # sign, and kept the hashtag texts.
- Negative words are useful for detecting negative emotion and issues. So we substituted words endingwith "n't" and other common negative words (e.g. nothing, never, none, cannot) as "negtoken".
- We removed all words that contain non-lettersymbols and punctuation. This included the removalof @ and http links. We also removed all theRTs.
- For repeating letters in words, our strategy wasthat when we detected two identical letters repeating,we kept both of them. If we detected morethan two identical letters repeating, we replacedthem with one letter. Therefore, "huuungryyy" and"sooo" were corrected to "hungry" and "so"."muuchh" was kept as "muuchh". Originally correctwords such as "too" and "sleep" were kept as they were.
- We used the Lemur information retrieval toolkitto remove the common stopwords. We keptwords like "much, more, all, always, still, only", because the tweets frequently use these words to expressextent. The Krovetz stemmer in the Lemurtoolkit was used to perform stemming in order tounify different forms of a word, such as pluralsand different forms of a verb.

4.2. Evaluation Measures for Multi-label Classifier

Commonly used measures to evaluate the performance ofclassification models include accuracy, precision, recall, and the harmonic average between precision and recall –the F1 score. For multi-label classification, the situation is slightly more complicated, because each document getsassigned multiple labels. Among these labels, some maybe correct, and others may be incorrect. Therefore, thereare usually two types of evaluation measures – examplebasedmeasures and label-based measures. Example based measures are calculated on each document (e.g.each tweet is a document, and also called an example here)and then averaged over all documents in the dataset,whereas label-based measures are calculated based oneach label (category) and then averaged over all labels(categories).

5. Comparison Experiment: Svm and M3l

SVM (Support Vector Machines) is one of the most usedand accurate classifiers in many machine learning tasks, butour 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-allheuristic. We applied weight of loss parameters that areproportional to the inverse of the percentages of tweets in or not in each category to account for the imbalanced categories.However, with the same training and testing datasets 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 alltweets.Then we applied the same training and testing datasets as above to an advanced SVM variation named Max-Margin Multi-Label (M3L) classifier. M3L is a state-of-theart multi-label classifier. Different from the oneversus-all heuristic, which assumes label independence,this classifier takes label correlation into consideration.We used the executable file of this algorithm provided bythe authors. The performance is better than the simplisticone-versus-all SVM classifier, but still not as goodas the Naive Bayes classifier. Table 4 and Fig. 3 show theevaluation measures using M3L. Because SVM is not aprobabilistic model, so Table 4 does not have probabilitythreshold values as Table 2 does.

6. Discussion, Limitations, and Future Work

This study explores the previously uninstrumented spaceon Twitter in order to understand engineering students'experiences, integrating both qualitative methods andlarge-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 itsuccessfully. Heavy study load leads to many consequencesincluding 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 conflictsand culture stereotypes existing among engineering students.Building on top of the qualitative insights, we implementedand evaluated a multi-label classifier to detectengineering student problems from Purdue University. This detector can be applied as a monitoring mechanismto identify at-risk students at a specific university in thelong run without repeating the manual work frequently.Our work is only the first step towards revealing actionableinsights from student-generated content on socialmedia in order to improve education quality. The classifieris designed to be a multi-label classifier in order to reconcilethis effect. If a tweet expresses correlation between"heavy study load" and "sleep problems", then it can becategorized into both categories. After all, any mathematicaland statistical models are simplification of real worldproblems to a certain extent. The comparison experimentwith M3L shows that this advanced model that accountsfor label correlation does not perform as well as the simpleNaïve Bayes model. Future work could specificallyaddress the correlations among these student problems.Finally, the workflow we proposed requires human effortfor data analysis and interpretation. This is necessarybecause our purpose is to achieve deeper understandingof the student experiences. To the best of our knowledge,there is currently no unsupervised automatic natural language processing technique that can achieve the depth ofunderstanding that we were able to achieve. There is atrade-off between the amount of human effort and thedepth of the understanding. The labels generated can beapplied to any similar datasets in other institutions todetect engineering student problems without extra humaneffort. Often times, manual analysis is timeconsumingnot only because of the time spent on analysing the actual data, but also the time spent on cleaning,organizing the data, and adapting the format to fit thealgorithms. We plan to build a tool based on the workflowproposed here combining social media data andpossibly student academic performance data. This toolcan assist in identification of students at risk. This toolwill provide a friendly user interface and integration betweenqualitative analysis and the classification and detectionalgorithms .Therefore, educators and researchersusing this tool can focus on the actual dataanalysis and investigate the types of learning issues that they perceive as critical to their institutions and students.This tool can also facilitate collaboration among researchersand educators on data analysis. Advanced naturallanguage processing techniques can be applied in the futureto provide topic recommendations and further augmentthe human analysis results, but cannot completelyrule out the human effort.Other possible future work could analyze students'generated content other than texts (e.g. images and videos),on social media sites other than Twitter (e.g. Facebook,Tumbler, and YouTube). Future work can also extendto students in other majors and other institutions.

7. Conclusion

Our study is beneficial to researchers in learning analytics,educational data mining, and learning technologies. Itprovides a workflow for analyzing social media data foreducational purposes that overcomes the major limitationsof both manual qualitative analysis and large scalecomputational analysis of user-generated textual content.Our study can inform educational administrators, practitionersand other relevant decision makers to gain furtherunderstanding of engineering students' college experiences.As an initial attempt to instrument the uncontrolledsocial media space, we propose many possible directionsfor future work for researchers who are interested in thisarea. We hope to see a proliferation of work in this area inthe near future. We advocate that great attention needs tobe paid to protect students' privacy when trying to providegood education and services to them.

8. 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. 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. Ya1939-1382 (c) 2013 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.suhara, and D. Lund, "Enabling engineering student success:The final report for the Center for the Advancement of EngineeringEducation," Morgan & Claypool Publishers, Center forthe Advancement of Engineering Education, 2010.