Mining Social Media Data for Understanding Students’ Learning Experiences

Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan

FALL 2015
COMP 5900X - Special Topics: Mining Software Repositories
Paper Presentation
By: Kenny Byrd
Outline

• Research Goals
• Overview
• Data Mining Workflow
• Step 1: Data Collection
• Steps 2-3: Inductive Content Analysis
• Step 4: Qualitative Results
• Step 5: Multi-label Classifier
• Step 6: Model Adaptation
• Step 7: Results
• Discussion, Limitations and Future Work
• Conclusion
Research Goals

The research goals of this study are:

1) to develop and demonstrate a workflow for making sense of social media data for educational purposes, and

2) to explore engineering students’ informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences.
Overview

• Huge amount of social media data (Facebook, Twitter, Youtube)

• These platforms allow students to share their experiences, vent emotion and stress, and seek social support

• Students discuss and share their everyday encounters in an informal and casual manner.
What does this mean?

• Well, compared to other approaches in Educational Data Mining (EDM) such as mining BlackBoard and other course management systems (CMS) for students’ academic performance,

• This study will focus on students’ social and emotional aspects based on their informal online conversations.

• This understanding can help with:
  • institutional decision-making on interventions for at-risk students
  • improvement of education quality
  • enhance student recruitment, retention, and success.
What was done?

• Conducted a qualitative analysis on samples taken from about 25,000 tweets related to engineering students’ college life.

• Found that engineering students encounter problems such as heavy study load, lack of social engagement, and sleep deprivation.

• Implemented a multi-label classification algorithm to classify tweets reflecting students’ problems into different categories.

• Then used the algorithm to classify students’ problems from about 35,000 tweets streamed at the geo-location of Purdue University.
Data Mining Workflow

Fig. 2. The width of gray arrows represents data volumes—wider indicates more data volume. Black arrows represent data analysis, computation, and results flow. The dashed arrows represent the parts that do not concern the central work of this paper. This workflow can be an iterative cycle.
Step 1: Data Collection

• Started by searching a social media monitoring tool Radian6 based on different combinations of keywords such as engineer, students, campus, class, homework, professor, and lab.

• The search logic grew complicated very quickly.

• During the month of November 2011, only 116 tweets relevant to college students were collected.

• A Twitter hashtag #engineeringProblems was found in this data set and occurred most frequently.
Step 1: Data Collection

• Using Radian6 again, they streamed tweets containing the hashtag #engineeringProblems over a period of 14 months and collected 25,284 tweets (from November 1st, 2011 to December 25th, 2012).

• For the second data set they used the Twitter Search API to collect tweets using the geo-location of Purdue University with a bounding box radius of 1.3 miles to cover the entire campus.

• From February 5th to April 17th, 2013 - 39,095 tweets were collected for the second data set.
Steps 2-3: Inductive Content Analysis

• Couldn’t find any algorithms to properly categorize the tweets 😞

• 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.

• And faulty assumptions are likely to arise if automatic algorithms are used without qualitative analysis.

• So, three researchers performed an inductive content analysis on the #engineeringProblems data set to manually analyze the Tweet body content.
Steps 2-3: Inductive Content Analysis

• Goals: identify what were the major worries, concerns, and issues that engineering students encountered in their study and life.

• Researcher A read a random sample of 2,000 tweets from the 19,799 unique tweets in the data set and developed 13 initial categories.

• Researcher A then sent all the categories with descriptions and samples to Researchers B and C.
Steps 2-3: Inductive Content Analysis

• All three researchers then discussed and collapsed the initial 13 categories into 5 prominent themes:
  1. heavy study load
  2. lack of social engagement
  3. negative emotion
  4. sleep problems
  5. diversity issues

• 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.
Prominent Themes

Fig.3. Number of tweets in each category of the 2,785 tweets analyzed. Note that a large number of tweets fall into “Others”, which indicates the “long tail” character of user-generated Internet content.
Step 4: Qualitative Results

• **Heavy Study Load:**

• Found that students express a very stressful experience in engineering. Not being able to manage the heavy study load leads to consequences such as lack of social engagement, lack of sleep, stress, depression, and some health problems.

• For example: “going to bed at 3 A.M. Still have about 8 hrs of homework and studying to do. . ..”
Step 4: Qualitative Results

• Lack of Social Engagement:
  • Found that students need to sacrifice the time for social engagement in order to do homework, and to prepare for classes and exams.

• For example: “Chemistry and calculus homework everyday of thanksgiving break”
Step 4: Qualitative Results

• Negative Emotion:

• Found that some tweets should be classified as “negative emotion” when it specifically expresses negative emotions such as hatred, anger, stress, sickness, depression, disappointment, and despair.

• For example: “is it bad that before I started studying for my tests today that I considered throwing myself in front of a moving car??”
Step 4: Qualitative Results

• **Sleep Problems:**

• Found that Students frequently suffer from lack of sleep and nightmares due to heavy study load and stress.

• For example: “I wake up from a nightmare where I didn’t finish my physics lab on time”
Step 4: Qualitative Results

• **Diversity Issues:**

• Found that students perceive a significant lack of females in engineering.

• This issue may be again related to the nerdy and anti-social image of engineering.

• For example: “‘Let's start with an example, tell me something you know nothing about’ – Professor . . . ‘girls.’ – Students. lol”
Step 4: Qualitative Results

• **Others: The Long Tail:**

A small number of common student problems appear in high frequency, and a large number of less common problems or noisy tweets each appear in very low frequency. This indicates a “long tail” character.

• **Examples:** curriculum problems, lack of motivation, procrastination, career and future worries, identity crisis, thought of switching majors, and physical health problems.
Step 5: Multi-label Classifier

• Built a multi-label classifier to classify tweets based on the 5 themes previously mentioned.

• Found the Naïve Bayes classifier to be very effective on the data set compared to other state-of-the-art multi-label classifiers.

• Pre-processed the texts before training the classifier to reduce noise in the text.
Step 5: Multi-label Classifier

• **Text Pre-Processing:**

• Removed all the #engineeringProblems hashtags. For other hashtags only the # sign was removed.

• Negative words that are useful for sentiment analysis (e.g., nothing, never, none, cannot) were substituted with “negtoken”.

• Removed all words that contain non-letter symbols and punctuation. This included the removal of @ and http links. Retweets were also removed.
Step 5: Multi-label Classifier

• Text Pre-Processing:

• For repeating letters in words, when two identical letters repeating were detected, both were kept. If more than two identical letters repeating were detected, they were replaced with one letter.

• Common stop words (most common words used in a language) such as “a, an, and, of, he, she, it” were removed to improve performance.

• Words were stemmed using the Krovetz stemmer in the Lemur toolkit:
  1. Transforming the plurals of a word to its singular form
  2. Converting the past tense of a word to its present tense
  3. Removing the suffix ‘ing’
Step 5: Multi-label Classifier

• Evaluation Measures:

• Commonly used measures to evaluate the performance of classification models include accuracy, precision, recall, and the harmonic average between precision and recall—the $F_1$ score.

• *Harmonic average*? Appropriate for situations when the average of rates is desired.

• For multi-label classification, the situation is slightly more complicated, because each document gets assigned multiple labels.
Step 5: Multi-label Classifier

• Classification Results:

• 70% of the 2,785 tweets were used for training (1,950 tweets), and 30 percent 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.
Step 5: Multi-label Classifier

From Table 2, we see that when the probability threshold value is 0.4, the performance is generally better than under other threshold values.

<table>
<thead>
<tr>
<th>$T$</th>
<th>Ex. $a$</th>
<th>Ex. $p$</th>
<th>Ex. $r$</th>
<th>Ex. $F_1$</th>
<th>Micro. $F_1$</th>
<th>Macro. $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1724</td>
<td>0.1724</td>
<td>1.0000</td>
<td>0.2931</td>
<td>0.2941</td>
<td>0.2551</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6223</td>
<td>0.6266</td>
<td>0.8188</td>
<td>0.6838</td>
<td>0.6536</td>
<td>0.5412</td>
</tr>
<tr>
<td>0.2</td>
<td>0.6630</td>
<td>0.6675</td>
<td>0.7850</td>
<td>0.7041</td>
<td>0.6841</td>
<td>0.5799</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6879</td>
<td>0.6934</td>
<td>0.7623</td>
<td>0.7142</td>
<td>0.7002</td>
<td>0.5960</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6996</td>
<td>0.7068</td>
<td>0.7463</td>
<td>0.7177</td>
<td>0.7070</td>
<td>0.6145</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7019</td>
<td>0.7091</td>
<td>0.7291</td>
<td>0.7135</td>
<td>0.7055</td>
<td>0.6116</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7052</td>
<td>0.7140</td>
<td>0.7205</td>
<td>0.7133</td>
<td>0.7061</td>
<td>0.6071</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7064</td>
<td>0.7152</td>
<td>0.7156</td>
<td>0.7125</td>
<td>0.7064</td>
<td>0.6054</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7060</td>
<td>0.7158</td>
<td>0.7113</td>
<td>0.7111</td>
<td>0.7054</td>
<td>0.6005</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7078</td>
<td>0.7176</td>
<td>0.7101</td>
<td>0.7119</td>
<td>0.7072</td>
<td>0.6025</td>
</tr>
<tr>
<td>1</td>
<td>0.7088</td>
<td>0.7199</td>
<td>0.7088</td>
<td>0.7125</td>
<td>0.7077</td>
<td>0.6028</td>
</tr>
<tr>
<td>Rand.</td>
<td>0.0412</td>
<td>0.0415</td>
<td>0.0424</td>
<td>0.0417</td>
<td>0.0391</td>
<td>0.0180</td>
</tr>
</tbody>
</table>

$T =$ probability threshold, Ex. $a =$ example-based accuracy (8), Ex. $p =$ example-based precision (9), Ex. $r =$ example-based recall (10), Ex. $F_1 =$ example-based $F_1$ (11), Micro. $F_1 =$ micro-averaged $F_1$ (16), Macro $F_1 =$ macro-averaged $F_1$ (17), Rand. = random guessing.
Step 5: Multi-label Classifier

• Used a random guessing program to first guess whether a tweet belongs to “others” based on the proportion this category takes in the training data set.

• If this tweet did not belong to “others”, it then proceeded to guess whether it fell into the rest of the categories also based the proportion each category takes in the rest categories.

• The random guessing program was repeated 100 times, and the average measures were obtained.
Step 5: Multi-label Classifier

### TABLE 3
Label-Based Accuracy and $F_1$ for Each Category Naïve Bayes versus Random Guessing

<table>
<thead>
<tr>
<th>Category</th>
<th>Label a.</th>
<th>Label $F_1$</th>
<th>Rand. a.</th>
<th>Rand. $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy Study Load</td>
<td>0.8514</td>
<td>0.4851</td>
<td>0.0756</td>
<td>0.0136</td>
</tr>
<tr>
<td>Lack of Social</td>
<td>0.9165</td>
<td>0.5072</td>
<td>0.0821</td>
<td>0.0099</td>
</tr>
<tr>
<td>Engagement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Emotion</td>
<td>0.9644</td>
<td>0.5039</td>
<td>0.0877</td>
<td>0.0062</td>
</tr>
<tr>
<td>Sleep Problems</td>
<td>0.9595</td>
<td>0.6916</td>
<td>0.0851</td>
<td>0.0077</td>
</tr>
<tr>
<td>Diversity Issues</td>
<td>0.9226</td>
<td>0.6882</td>
<td>0.0837</td>
<td>0.0088</td>
</tr>
<tr>
<td>Others</td>
<td>0.7432</td>
<td>0.8107</td>
<td>0.0527</td>
<td>0.0617</td>
</tr>
</tbody>
</table>

Label a. = label-based accuracy (12), Label $F_1$ = label-based $F_1$ (15),
Rand. a. = random guessing label-based accuracy,
Rand. $F_1$ = random guessing label-based $F_1$.

Table 3 shows the label-based accuracy and $F_1$ measure for each of the six categories when $T = 0.4$ compared with random guessing.
Step 5: Multi-label Classifier

• Concluded that:

• The Naïve Bayes classifier has not only achieved significant improvement from the random guessing baseline, but also exceeded the performance of state-of-the-art multi-label classifiers on the data set.
Step 6: Model Adaptation

• The Naïve Bayes multi-label classifier was used to detect engineering student problems from the Purdue data set.

• Took a random sample of 1,000 tweets from the Purdue data set, and found that no more than 5% of these tweets were discussing engineering problems.

• To make the training set better adapt to the Purdue data set, a random sample of 5,000 tweets from the Purdue data set was added to the 2,785 #engineeringProblems tweets, and was labeled as “others”.
Step 7: Results

• Has the ability to detect potential student problems from tweets.

• Not making the claim that these users shown are definitely at-risk students, since most of these users only posted less than 10 percent of problems among all tweets they have posted.

• The trained detector can be applied as a monitoring mechanism in the long run to identify severe cases of at-risk students.

• For example, a future student may post a large number of tweets and more than 90% of them are about study problems or negative emotions.
Step 7: Results

Fig. 4. Number of tweets for each issue detected from the Purdue tweet collection. We do not discuss “others”, because we detect tweets reflecting these five problems from large number of Purdue tweets.
Step 7: Results

Fig. 5. Top 15 users in the Purdue tweet collection who posted the most on the five engineering problems.
### Step 7: Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Top 25 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy Study Load</td>
<td>hour, homework, exam, day, class, work, negtoken, problem, study, week, toomuch, all, lab, still, out, time, page, library, spend, today, long, school, due, engineer, already</td>
</tr>
<tr>
<td>Lack of Social Engagement</td>
<td>negtoken, Friday, homework, out, study, work, weekend, life, class, engineer, exam, drink, break, Saturday, people, social, lab, spend, tonight, watch, game, miss, party, sunny, beautiful</td>
</tr>
<tr>
<td>Negative Emotion</td>
<td>hate, f***, shit, exam, negtoken, week, class, hell, engineer, suck, study, hour, homework, time, equate, FML, lab, sad, bad, day, feel, tired, damn, death, hard</td>
</tr>
<tr>
<td>Sleep Problems</td>
<td>sleep, hour, night, negtoken, bed, allnight, exam, homework, nap, coffee, time, study, more, work, class, dream, ladyengineer, late, week, day, long, morning, wake, awake, no-sleep</td>
</tr>
<tr>
<td>Diversity Issues</td>
<td>girl, class, only, negtoken, guy, engineer, Asia, professor, speak, English, female, hot, kid, more, toomuch, walk, people, teach, understand, chick, China, foreign, out, white, black</td>
</tr>
</tbody>
</table>
Discussion, Limitations and Future Work

• 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.

• Also, students’ awareness of identity management online may increase overtime.

• The “manipulation” of personal image online may need to be taken into considerations in future work.

• The fact that the most relevant data found on engineering students’ learning experiences involve complaints, issues, and problems does not mean there is no positive aspects in students’ learning experiences.
Discussion, Limitations and Future Work

• Future work can compare both the good and bad things to investigate the trade-offs with which students struggle.

• Students tend to complain about issues and problems on social media. This may imply that social media serve as a good venue for students to vent negative emotions and seek social support.

• Therefore, future work can be done on why and how students seek social support on social media.

• Future work can also extend to students in other majors and other institutions.
Conclusion

• The research 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.

• The study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students’ college experiences.
Thank you
Questions

1. Has anyone taken any NLP courses? What other text pre-processing would you do?

2. Do you think another prominent theme could have emerged from the “Others” category and should have been used as well?

3. Do you think the time period in which the tweets collected would make much of a difference? #engineeringProblems data set was from November 1st, 2011 to December 25th, 2012 and Purdue data set was from February 5th to April 7th, 2013 maybe compared with a data set collected now.

4. How would you extend this research?
The End!