



Enhancing Sentiment Classification using Multi-label Text Mining

B.Narendra and T. Kavitha

Asst. Professors,
Sree Vidyanikethan Engineering College, A. Rangampet,
Tirupati, AP, INDIA

Abstract: Now-a-days there are many social networking sites like Twitter, Facebook. Students post their opinions and feelings through these sites, because these sites provide a way to present their informal conversations. The data from these sites provide a useful knowledge to know the student behaviour. But the problem here is the large volume of data. It is difficult for an automatic algorithm for deeper understanding of the dataset whereas manual checking takes more time to analyse the large amount of data. Here we focussed on the Engineering students to know their educational experiences.

Keywords: Text Pre-processing, Wordnet Interface, Parsing, Hash Tags

I. INTRODUCTION

Students post their feelings, emotions, opinions through the social networking sites like Facebook, Twitter and seek their social support from their friends. These digital footprints provide a greater knowledge for the researchers to understand the student's feelings, opinions, and emotions. The researchers analyse these data and understand the student's problems regarding the education. These analysis results help to make necessary decisions to improve the educational system so that students can be relieved from the heavy burden.

Previously the educational researchers organise the surveys, meetings to know the students feedback regarding the education. But the problem here is, it is difficult to collect the feedback from the large number of students and takes more time to finalise the feedback from those feedback.

Later some more advancement has taken over and educational researchers instead of conducting surveys, they tried to collect the feedback through online [1], [2], [3]. This reduces the burden to some extent and quickly analyse the results.

The research goals of this paper is

- To develop a workflow that analyse the both qualitative analysis and huge amount of the data.
- To analyse the student's informal conversations and make a decision based on these.

In this paper we analysed the data from the Twitter because in Twitter we use only 140 character to post tweets and it is easy to collect the data through the Hashtags. Here we used the #engineering Problemshashtag.

These analyses can make to improve the educational quality and achieve the success [4] in the educational system.

II. RELATED WORK

A. Mining Twitter Data:

From the past researchers have analysed the social media content for the various purposes like Healthcare [5], marketing [6], athletics [7].

One of the most important aspects of this theory is the notion of both the 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 and opinions. Hence the back-stage data helps major role in our study.

Before we do review studies on Twitter from the fields of data mining, machine learning, and natural language processing. They cover a wide range of topics including information propagation and diffusion [8], popularity prediction [9], event detection [10], topic discovery [11], and tweet classification [12], to name a few. Amongst these topics, tweet classification is most relevant to our study.

B. Data Collection:

In twitter we collected the data using the keywords like engineer, students, campus, class, homework, professor, and lab. The Twitter APIs [13] can also be configured to accomplish this task. In our paper we collected the data from the #engineeringProblemshashtag.

III. INDUCTIVE CONTENT ANALYSIS

Researchers on analysing the tweets classified the students problems into different categories like :heavy study load, lack of social engagement, negativeemotion, sleep problems, and diversity issues.

A. Heavy Study Load:

This category comes under classes, homework, exams, and labs dominate the students' life. Libraries, labs, and the engineering building are their most frequently visited places.

Example

"Study over 30 hours for a test", "so much homework, so little time",

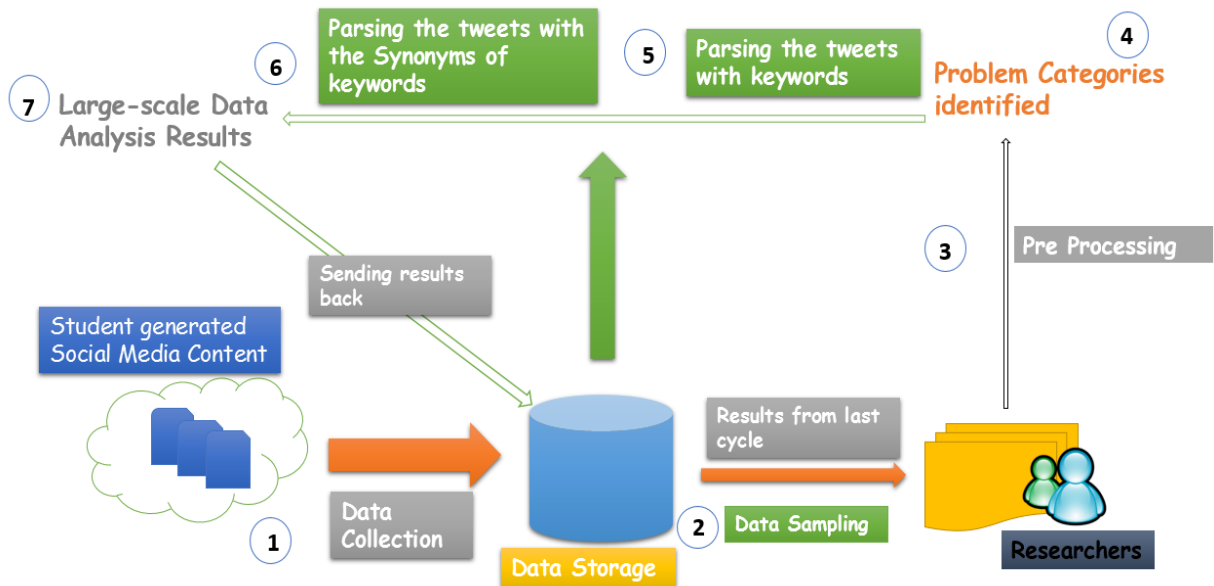


Figure: 1

Figure. 1. The workflow we developed for making sense of social media data integrates qualitative analysis and data mining techniques. The width of grey arrows represents data volumes—wider indicates more data volume. Black arrows represent data analysis, computation, and results flow. This process can occur iteratively.

B. Lack of Social Engagement:

The study show that students need to sacrifice the time for social engagement in order to do homework, and to prepare for classes and exams.

Example,

“I feel like I’m hidden from the world—life of an Engineering student”.

C. Negative Emotion:

There are a lot of negative emotions present in the hash tag #engineeringProblems has a negative emotion. We only categorize a tweet as “negative emotion”.

Example

“Is it bad that before i started studying for my tests today that i considered throwing myself in front of a moving

D. Sleep Problems:

Sleep problems are widely common among engineering students.

Example

“I won’t sleep for the next three days because of heavy load and stress”, kept both of them. If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, “huuungryyy” and “sooo” were corrected to “hungry” and “so”. “muuchh” was kept as “muuchh”. Originally correct words such as “too” and “sleep” were kept as they were.

E. Diversity Issues:

Some studies suggest that students perceive a significant lack of females in engineering.

Example

“eighty five kids leaving the classroom before mine. Of those 85, four are girls.

IV. ANALYSING TWITTER DATA

A. Text Pre-processing:

- We removed all the #engineeringProblems hashtags. For other co-occurring hashtags, only remove the # sign.
- Negative words are useful for detecting negative emotion and issues. So substitute words endingn’t (haven’t,) with negative words.
- We removed all words that contain non-letter symbols and punctuation. This included the removal of @ , http links and also all the RTs.
- For repeating letters in words, our strategy was that when we detected two identical letters repeating, we The synonyms for a particular keyword can be retrieved by using WordNet database files. These are of open source.

B. Parsing Twitter Data:

- Read the each tweet from the pre-processed twitter data set which is free from the noise.
- Compare the each word in the tweet with the keywords shown in the table 1.
- If the any word match with given keywords of the particular category, the particular category is incremented once.
- Sometimes one tweet may consist of 2 or more keywords of different categories. (For example the tweet “Just downed a giant cup of coffee , now it’s

time for that load of homework I put off all weekend #engineeringproblems “ tweet contain 1 Heavy Study Load, 1 Lack of Social Engagement and 1 Sleep Problems)In such a situation each category problem should be incremented by 1.

- e. Similarly parse all the tweets with the given keywords of different categories.
- f. While parsing the each tweet should maintain the count of the keywords, so that the particular category should be incremented to the count value of the keywords.
- g. Finally the category which has the highest count can be treated as most of the students are suffering from that problem(may HeavyStudyLoad/Lack of Social Engagement / Negative Emotion / Sleep Problems / Diversity Issues).
- h. Once the problem category was identified necessary steps should be taken by the Educational Institutions to reduce the burden of students from that problem.

C. Parsing with Synonyms:

To improve the accuracy of our model, besides parsing with the given keywords, also parse the tweets with the synonyms of the keywords(For example consider the tweet “DLD exam preparation will be the death of me” Let the keyword “exam” belongs to the HeavyStudyLoad Category. Some students may tweet (DLD examination preparation will be the death of me) or (DLD test preparation will be the death of me). If we consider above 3 tweets, Examination and Test are the synonyms of the word Exam. So the below two tweets also come under the HeavyStudyLoad category. Hence by incorporating the synonyms in our model, it improves the accuracy of model to the greater extent.

D. WordNet Interface:

An interface to WordNet using the Jawbone Java API to WordNet. (<http://wordnet.princeton.edu/>) is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing aon conceptual-semantic and lexical relations. Please note that WordNet(R) is a registered trade name WordNet available to research and commercial users free of charge provided the terms of their license (<<http://wordnet.princeton.edu/wordnet/license/>>) are followed, and proper reference is made to the project using an appropriate citation <http://wordnet.princeton.edu/wordnet/citing-wordnet/>

E. Algorithm Implementation:

Step1: Collect the dataset using the hashtag#engineeringProblems .

Step2: Pre-process the dataset

2.1: Remove the hash tags.

2.2: Remove the negtokens

2.3: Remove the unnecessary punctuation like http links, @ symbols, www symbols.

2.4: Remove the spelling mistakes present in the Tweets so that parsing becomes easier.

Step3: Parse the data set with the keywords present in the Table I

Step4: Parse the data set with the synonyms of the Keywords.

Step5: Identify the problem category

Step6: Find the performance evaluation measures.

Table I

CATEGORY	KEYWORDS
Heavy study-load	hour, homework, exam, day, class, work, negtoken, Problem, study, week, too much, all, lab, still, out, Time, page, library, spend, today, long, school, due, Engineer, already.
Lack of Social-Engagement	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, all.
Negative Emotion	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.
Sleep Problems	sleep, hour, night, negtoken, bed, all night, exam, homework, nap, coffee, time, study, more, work, class, dream, lady engineer, late, week, day, long, morning, wake, awake, no sleep.
Diversity Issues	girl, class, only, negtoken, guy, engineer, Asia, professor, speak, English, female, hot, kid, male, too much, walk, people, teach, understand, chick, china, foreign, out, white, black.
Precision	38/70 = 0.54
Recall	38/32 = 1.1875
F_Score	(0.54*1.1875)/((0.54+1.1875)/2) = 0.5538
Precision	40/70 = 0.5714
Recall	40/30 = 1.3333
F_Score	(0.5714*1.3333)/((0.5714+1.3333)/2) = 0.7255

V. PERFORMANCE EVALUATION

Without Synonyms

$$\text{Precision} = \frac{\{\text{Relevant} \cap \text{Retrieved}\}}{\{\text{Retrieved}\}}$$

$$\text{Recall} = \frac{\{\text{Relevant} \cap \text{Retrieved}\}}{\{\text{Relevant}\}}$$

$$\text{F_Score} = \frac{\text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})/2}$$

With Synonyms

Here **Relevant** means the count of the tweets that have included more than one category problem in it.

Retrieved = Total number of tweets in the dataset

Relevant = The count of tweets that have only one problem category included in it.

VI. LIMITATIONS AND FUTURE WORK

In our paper we chose to focus on the problems that could be the most informative for improvement of education quality. Future work can compare both the good and bad things to investigate the tradeoffs with which students struggle. From another aspect, 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 why and how students seek social support on social media sites.

VII. CONCLUSION

Our study is beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analysing 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.

VIII. REFERENCES

- [1]. R. Ferguson, "The State of Learning Analytics in 2012: A Review and Future Challenges," Technical Report KMI-2012-01, Knowledge Media Inst. 2012.
- [2]. R. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," J. Educational Data Mining, vol. 1, no. 1, pp. 3-17, 1, pp. 7-15, 1997.
- [3]. S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez, "Microblogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging-Supported Classroom," IEEE Trans.
- [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]. S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Law-son, "'I Can't Get No Sleep': Discussing #Insomnia on Twitter," Proc. ACM Ann. Conf. Human Factors in Computing Systems, pp. 1501-1510, 2012.
- [6]. M.J. Culnan, P.J. McHugh, and J.I. Zubillaga, "How Large US Companies Can Use Twitter and Other Social Media to Gain Business Value," MIS Quarterly Executive, vol. 9, no. 4, pp. 243-259, 2010.
- [7]. M.E. Hambrick, J.M. Simmons, G.P. Greenhalgh, and T.C. Greenwell, "Understanding Professional Athletes' Use of Twitter: A Content Analysis of Athlete Tweets," Int'l J. Sport Comm., vol. 3, no. 4, pp. 454-471, 2010.
- [8]. D.M. Romero, B. Meeder, and J. Kleinberg, "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, 2009. Learning Technologies, vol. 4, no. 4, pp. 292-300, Oct.-Dec. 2011.
- [9]. R. Bandari, S. Asur, and B.A. Huberman, "The Pulse of News in Social Media: Forecasting Popularity," Proc. Int'l AAAI Conf. Weblogs and Social Media (ICWSM), 2012.
- [10]. T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," Proc. 19th Int'l Conf. World Wide Web, pp. 851-860, 2010.
- [11]. W. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," Proc. 33rd European Conf. Advances in Information Retrieval, pp. 338-349, 2011.
- [12]. D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," Proc. 23rd Int'l Conf.
- [13]. "Using the Twitter Search API | Twitter Developers," <https://dev.twitter.com/docs/using-search>, 2013.