ISSN: 2321-7782 (Online)

Volume 3, Issue 2, February 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Understanding Students' Academic Experiences through Mining Social Media Data

Shashank Pawar[1]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Chinmay Phadkule[2]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Sayali Shinde[3]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Kunal Patil[4]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Abstract: Students' informal conversations on social media (e.g. Twitter, Facebook) shed light into their educational experiences. Such uninstrumented environments can provide valuable knowledge to inform student learning. However this task can be challenging. Human interpretation is required in dealing with social media. However, the growing scale of data demands automatic data analysis techniques. In this project, we shall develop a workflow to integrate both qualitative analysis and large-scale data mining techniques. We shall collect and analyse students' Twitter posts to understand issues and problems in their educational experiences. Based on these results, we shall implement a multi-label classification algorithm to classify tweets reflecting students' problems. We then will use the algorithm to train a detector of student problems from the tweets streamed. This work, for the first time, presents a practical methodology and results that show how informal social media data can provide insights into students' experiences.

Keywords: Twitter, Naïve Bayes, API, Social Media, Dataset.

I. Introduction

Social media plays important role in today's era of information and technology. It also plays crucial role in understanding students' views on academic curriculum, atmosphere of collages and study related sentiments. Students share their joys and sorrows related to studies on social media in the form of judgmental comments, tweets, posts etc. One important reason why social media can be relayed on is that the comments and posts are spontaneous emotions of students. They are not much thought over as we usually often do while answering surveys. Students share their daily experiences and happenings of daily classes in a very casual and informal way.

These informal footprints of students can be very useful and reliable for researchers to understand student's learning behavior outside the classroom. These studies can be very useful and may prove revolutionary for an educational institute as crucial changes can be made in educational nature of the institute. Different cultures of education can be cross checked and fused together in order to make our educational system sustainable, student friendly, global.

II. Research Goal And Educational Domain

Major goals of our research study are:

1) To make the bulk amount of data sense making for educational purposes.

2) To make friendship between large scale data mining techniques and qualitative analysis of the data.

3) To explore engineering students informal tweets on twitter in order to analyse the issues and problems faced by engineering students.

We selected engineering students' problems for our research purpose.

The major reasons were:

1) We ourselves are engineers and know the problems engineers face in engineering schools.

2) Collages face problems with students' recruitments and retention issues.

3) Engineers (IT industry) partly helps in growth of GDP of nation as well as they are crucial work-force and think tank of a nation. So their academic problems must be tackled.

4) Based on the results of our paper educational institutions can make changes in their policies.

5) Twitter is a popular social media site. Its matter is very concise (no more than 140 characters per tweet). Twitter provides free APIs that can be used to stream data.

It was very important for us to analyse the quality of the data (tweets) not ignoring the huge quantity of data that was/is continuously being put on twitter. So we needed to bridge and integrate method of quality research and large scale data mining techniques. The bridge we created was somewhat like this.

In order to get better understanding of data we used qualitative data mining algorithm based on human interpretation. We apply algorithm to another large scale data and unexplored dataset so that manual methods are augmented. We kept refining the model based on further human feedback.

III. System Architecture

Now**,** we see how actually the architecture will work in real. We are going to create an computational flow of the architecture. It will help us in designing the model with ease. Imagine an application wherein all the problems of students are present and all what our institutions need to do is to take actions. No special surveys and no thousands of feedback forms.

We divided the entire project into six modules. First step is to access the web page of twitter through HTML startup link. In this way we will be able to extract all the HTML code from twitter page. In second step we have two ways to get data for analysis one is to use API to get the tweets and select the area and Google server finds the co-ordinates of the area. All the tweets in the respective area will be analyzed and educational tweets will be sorted out based on the data sets we provide to our system and other way is to extracted html code is used to find user comments and to preprocess those comments and find useful comments (using predefined dataset). Preprocessing includes stop word removal, stemming, etc. We are going to provide stop words file to this module.

The useful comments found in earlier step will be used in this step to find user's polarity about the statement (positive or negative). We are going to use positive and negative words dataset for this module.

The user comment and its polarity will be processed and the grouping of comments into different domains will be done in this next step. Various grouping parameters we are going to consider here such as according to polarity, according to domain, etc. Categorization will take place using classification algorithm and we will get the results in the form of students' problems distributed in various domains.

IV. Collection of Data

Collection of data is a tedious task to collect social media data related engineering students' academic problems due to irregularity and assortment of language. Collection of twitter data is important in order to analyse it and perform a research study of the students' problems. The data collection method used by us comprises of our collage twitter page and related links on twitter so as to gain wide access to the database of twitter. We used the readymade data sets of some researches to get detailed and most probable results. There is lot of noise (unwanted information) present in Twitter posts or tweets. A group of words which we call data sets are used to remove noise. Preprocessing also removes noise. From limited number of relevant comments we found Twitter hashtag #Engineering Problems most useful in order to get engineering students' twits. #ladyEngineer,#engineeringMajors,#switch-ingMajors,#collegeProblems were some other hashtags which were found important in order to retrieve data.

V. Removal of Noise

The data we collected from twitter had large amount of noise in it. It was necessary to remove the unwanted words, letters, character and hashtags in order to analyse the content for classification purpose. So data we extracted from twitter was preprocessed as-

1) Students are in a mostly casual mood while posting on twitter. So often they type certain words which do not fit in rules of English language. The words like huuuuungry ,sooorrry etc are some words which can be understood only by human interpretation. So we decided to if we decided that if more than two identical letters repeating are there, we replace them with one letter. But this method had its own drawbacks.

2) Removal of all the #engineering Problems hashtags. For other co-occurring hashtags, we removed the # sign, and kept the hashtag texts.

3) We used negative tokens to detect negative emotions.

4) We removed all words that contain non-letter symbols and punctuation. This included the removal of @ and http links. We also removed all the RTs.

These were some of the preprocessing that we performed on the data derived from twitter.

VI. Categories

We studied and analyzed the data and concluded that most of the comments fell into some repeated categories. These categories are given below-

Curriculum issues, heavy load of study, study difficulties, imbalanced life, future and carrier worries, lack of gender diversity, sleep problems, stress, lack of motivation, physical health problems, nerdy culture, and identity crisis.

These were the categories wherein most of the comments fell. Among these categories also we found certain categories more important for our research because there most of the comments were from these five categories. So we decided to focus on these five major categories. They were- Heavy study load, lack of social evolvement, negative feelings, sleep problems and diversity of place, culture and stratum.

VII. Classification Algorithm

Multi label classifier can be implemented by converting it into multiple single label classifiers. One of the simple methods is one versus all also known as binary relevance. The multipliable classifier we used to classify the data was Naïve bayes classifier.

VIII. Limitations

1) We implemented and evaluated a multi-label classifier to detect engineering student problems from JSPM page on twitter and other related pages. So we will not be able to find the problems of overall students.

2) All students are not active on twitter, so we may only finds the ones who are more active and more likely to exposed their thoughts and feelings.

3) The fact that most relevant data we found on engineering students' experiences involve complaints. Positive side is neglected because we want to emphasis on students' problems

4) We only selected most occurring top five themes, other themes were boycotted for reason of ease.

5) Emotions detection is not a part of our project.

IX. Future Work

Our work is only the first step towards revealing actionable insights from students-generated content on social media in order to improved educational quality.

The "manipulation" of personal image online may need to be taken into considerations in future work. Future work can be done on why and how students seek social support on social media sites. Future work can be done to design more sophisticated algorithms in order to reveal the hidden information. Other possible future work could analyze student's generated content other than texts (images and videos) on social media sites other than twitter. Future work can also extend to students in other majors and other institutions.

X. Conclusion

Our study is beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a flow to analyse 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 institutions 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. We 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.

Acknowledgement

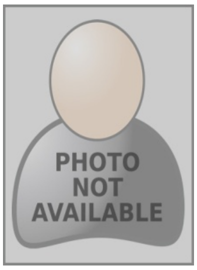
We would like to express our gratitude towards **Prof.H.A.Hingoliwala** whose support and consideration has been an invaluable asset during the course of this project. First and foremost, we would like to thank our guide **Prof.S.R.Ghule** for providing us with their invaluable guidance throughout the course of this project. It would have been an almost impossible task to complete this project without his support, motivation, valuable suggestions and criticism. We convey our gratitude to our respected **Head of Department, Prof.H.A.Hingoliwala** for his motivation, guidance, and criticism and also for providing various facilities, which helped us greatly in the course of this project.

And, last but not the least we would like to thank **Principal M.G.Jadhav** and all the teaching, non-teaching staff of Computer Department and also our friends for directly or indirectly helping us for the project completion and all the resources provided.

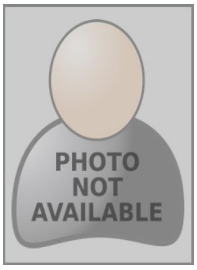
References

1. Mining social media data for understanding students learning process.
2. Social networks exploration for educational data mining by Pedro Miguel Terras Crespo.
3. Mining Social Media: A Brief Introduction Pritam Gundecha and Huan Liu A Tutorial in Operations Research INFORMS 2012.
4. G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," Educause Review, vol. 46, no. 5, pp. 30– 32.

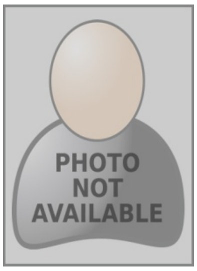
AUTHOR(S) PROFILE



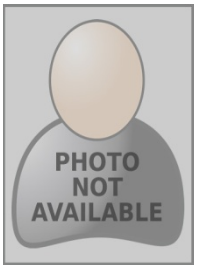
Mr. Shashank Pawar currently BE student in the Computer Science from Jayawantrao Sawant College of Engineering. And my research interested areas are in the field of Java, Data mining and web development.



Mr. Chinmay Phadkule currently BE student in the Computer Science from Jayawantrao Sawant College of Engineering. And my research interested areas are in the field of Java, Data mining and web development.



Ms. Sayali Shinde currently BE student in the Computer Science from Jayawantrao Sawant College of Engineering. And my research interested areas are in the field of java, Network Security, and Data mining.



Mr. Kunal Patil currently BE student in the Computer Science from Jayawantrao Sawant College of Engineering. And my research interested areas are in the field of Java, Data mining and web development.