Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.704 – 709

RESEARCH ARTICLE

Student Behaviour Predictions using Social Media Network in Hadoop Framework

Vaibhav Wadhai[1], Suhaib M. Ansari[2], Saikiran Chepuri[3]

[1]SITE & VIT University, India
[2]SITE & VIT University, India
[3]SITE & VIT University, India
[1] vibs.wadhai@gmail.com; [2] es_em31@yahoo.co.in; [3] saikiranchepuri35@gmail.com

***Abstract*: *Now a day's social media provide a range of opportunities for understanding human behaviour through the large aggregate data sets that their operation collects. Data Mining is very useful in the field of education especially when examining students' learning behaviour in online learning environment. So Student's casual conversations on social media (Twitter) shed light into their educational experiences like opinions, feelings, and concerns about the learning process. Data from such instrumented environments can provide valuable knowledge to inform student learning. Analysing 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' tweets to understand issues and problems in their educational experiences. We found engineering students encounter problems such as heavy study load, lack of social engagement, negative emotions and sleep deprivation. So we implemented Range classification algorithm to classify tweets reflecting students' problems. It can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' college experiences*.**

Keywords: social networking, web text analysis, computers and education, Twitter, Map reduce

I. INTRODUCTION

Social media places such as Twitter, Facebook, and YouTube offer great sites for students to share happiness and struggle, escape feeling and stress, and pursue social support. On various social media places, students discuss and share their everyday meetings in an informal and casual manner. Students' numerical paths provide vast whole of understood knowledge and a whole novel viewpoint for educational investigators and practitioners to understand students' involvements external the precise classroom environment. This considerate can inform official decision making on involvements for at risk students, upgrading of education quality, and thus improve student recruitment, retention, and success [1]. The plenty of social media data offers chances to know students' experiences, but also increases procedural difficulties in making logic of social media data for enlightening purposes. Just visualize the pure data volumes, the diversity of Internet slangs, the randomness of locations, and timing of students posting on the web, as well as the difficulty of students' experiences. Uncontaminated manual analysis cannot contract with the ever rising regulation of data, while clean automatic algorithms typically cannot capture in complexity meaning inside the data [2].

Usually, educational investigators have been using approaches like surveys, interviews, focus groups, class classroom actions to gather data associated to students' learning experiences. These approaches are usually very time consuming, thus cannot be copied or repeated with high occurrence. The measure of such trainings is also usually limited. When prompted about their involvements, students need to replicate on what they were thinking and doing in the past that may have become hidden over time.

The research goals of this study are I) to govern a workflow of social media data sense making for educational purposes, integrating both qualitative analysis and large scale data mining techniques and II) to discover engineering students' casual

discussions on Twitter, in demand to understand matters and difficulties students meeting in their information experiences. Built on considerate of concerns and problems in students' life, officials and educationalists can make extra knowledgeable choices on proper involvements and facilities that can support students overwhelmed walls in learning.

II. LITERATURE SURVEY

The hypothetical basis for the cost of casual data on the net can be located drawn from Goffmans philosophy of social performance [3]. Although established to describe face to face interactions, Goffmans philosophy of social performance is extensively used to explain decided connections on the web today [4]. One of the greatest essential features of this philosophy is the notion of front-stage and back-stage of people's social performances. Compared with the front stage, the peaceful atmosphere of back stage typically inspires more impulsive actions. Whether a social situation is front stage and back stage is a comparative matter.

Many educations demonstration that social television users may decisively manage their connected identity to 'look better' than in actual life. Extra studies show that around is a absence of alertness about handling online identity between college students, and that new people typically favour social media as their private space to fall out with peers outside the vision of parents and teachers. Students online chats disclose features of their involvements that are not simply understood in official classroom settings, thus are typically not recognized in scholastic literature. The adequately of social media data gives chances but also offerings technical problems for learning large rule casual written data.

Mining Twitter Data:

Researchers have studied Twitter content to generate precise information for their particular topic domains. Gaffney [5] examines tweets with hashtag '#iranElection' by histograms, employer networks, and occurrences of top keywords to enumerate online involvement. Like educations have been showed in extra fields plus healthcare [6], marketing [7], and athletics [8]. Study methods used in these studies typically include qualitative content study, linguistic study, network study, and some naive approaches such as word clouds and histograms. In our study, we constructed a classification model created on inductive content analysis. This model was then useful and confirmed on a make new dataset. So, we highlight not only the understandings gained from one data-set, but also the request of the classification algorithm to additional datasets for noticing student problems. The human strength is thus augmented with large scale data study.

I review studies on Twitter from the areas of data mining, machine learning, and natural language processing. These readings typically have more prominence on statistical models and procedures. They cover extensive variety of themes with information circulation and diffusion, approval prediction, event detection, topic detection, and tweet classification [9], [10], [11], [12]. Tweet classification is related to our study.

In starting, binary classification and multi class classification algorithm are used based on number of classes. Binary classification represents only two classes and multi class classification represents more than two classes. These are single-label classification systems. Each data point can only fall into one class single label classification here classes are mutually exclusive. Multi label classification allows each data point to fall into some classes at the same time.

Most studies on tweet classification are binary classification on relevant and irrelevant content [9], and multi class classification on classes like news, events, opinions, deals, and private messages [10]. Sentiment analysis is very standard three class classification on positive, negative, or neutral emotions/opinions [11]. Sentiment analysis is most useful for mining customer sentiments on products and companies over their appraisals or online posts. It finds extensive acceptance in marketing and customer relationship management (CRM).

In our study, I implemented range classification where I find out the sentiment value of every tweet and find out on which problems category it falls.

III. DATA COLLECTION

It is difficult to collect social media data related to students practises because of the irregularity and variety of the language used .I found a Twitter hashtag #CollegeProblems happening most often. Students used the hashtag #engineeringProblems to post about their opinions of being engineering majors. This was the popular hashtag precise to engineering student's college life. Also I identified several less popular but require 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.

Existing System:

Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, class room activities to collect data related to students learning experiences. When encouraged about their experiences, students need to reflect on what they were thinking and doing in the past which may have become unnoticed. These methods are typically very time consuming, and thus cannot be duplicated or repeated with high frequency.

Proposed work:

We will connecting social networks to do the analysis of the student's activity .Twitter tweets of Students have huge of data which can gives numerous activity of them. By getting the information we can predicate the activity of the students. I focused on engineering students tweets to understand concerns and problems in their educational experiences. This understanding can inform institutional decision making on interventions for at risk students, upgrading of education quality, and thus improve student enrolment, retention, and success.

Development of Categories:

Heavy Study Load:

I found that, homework, classes, exams, and labs control the student's life. Libraries, labs, and [15] the engineering building are their most frequently visited places. Some tweets are "Study over 30 hours for a test", "so much homework, so little time", and "C++ CAE project due Tuesday, Mfg project Wednesday, 25 PageTech Repot Wednesday + heavy homework load.

Lack of Social Engagement:

It shows that students need to expense the time for social meeting in order to do homework, and to make for classes and exams. For example, "I feel like I'm hidden from the world - life of an Engineering student".

Negative Emotion:

It precisely express negative emotions such as hatred, anger, stress, sickness, depression, disappointment, and despair.

Sleep Problems:

It finds that sleep problems are generally common between engineering students. Students regularly suffer from absence of sleep and imaginings due to heavy study load and stress. For example, "Napping in the common room because I know I won't sleep for the next three days".

Others: the Long Tail:

A large number of tweets fall below this category. Many tweets in this category do not have a strong meaning. Other tweets in this category do replicate various concerns that engineering students have but seen in actual small volumes.

IV. ARCHITECTURE

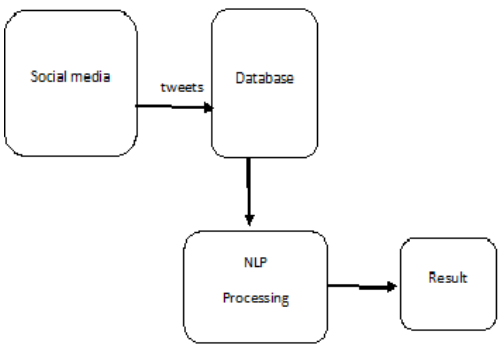

Fig1: Proposed Architecture

From the social media data like twitter, Facebook, YouTube we can share lot of data. Here I am taking required tweets from twitter and uploaded into hadoop data file system. Then applied NLP processing to it.

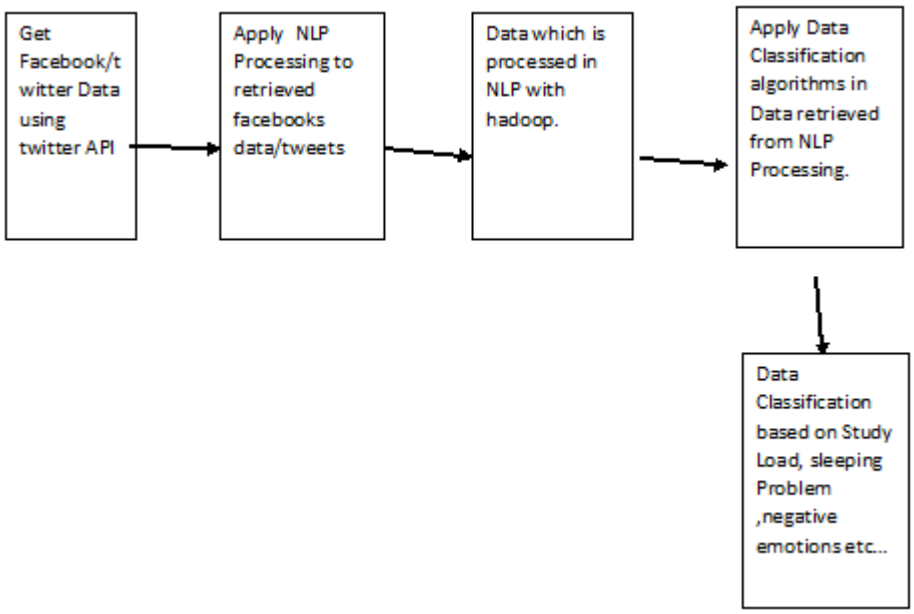

Fig2: Flow of data:

V. ALGORITHM

Natural Language Processing (NLP):
It is used to find nouns, adverb, adjectives and keyword in text.

- The process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation.
- The field of NLP is primarily concerned with getting computers to perform useful and interesting tasks with human languages.
- The field of NLP is secondarily concerned with helping us come to a better understanding of human language.

Classification Algorithm:
For each category mentioned in the above section, find out the sentiment of the each category using sentiment analysis (sentiword.net).Then made the intervals between the classes using range of sentiment value. And apply the words obtained in nlp processing to the classification.

VI. EVALUATION AND RESULT

I have retrieved tweets from twitter (social media). Then I used NLP processing and sentiment analysis. After getting output from NLP Process I used range classification on sentiment value of problem category and analysed the result.

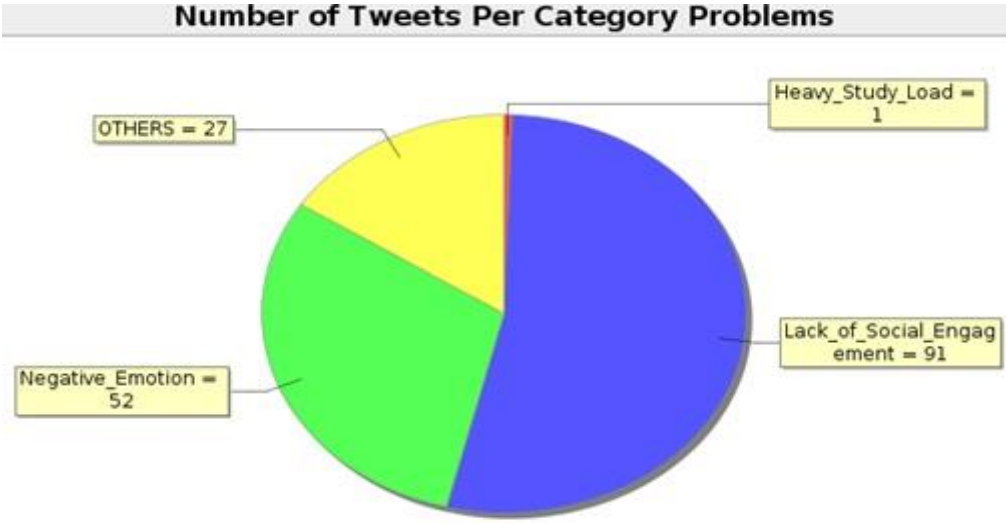

Fig3:Classification of tweets:

In the above fig, you can see that, tweets are classified into four categories such as heavy study load, negative emotions, lack of social engagement and others. If the tweet does not fall into any of the mentioned category then it fall into others category. This work has been done in hadoop framework. From above fig we can say that student having lack of social engagement problem. Hadoop framework uses mapper and reducer functions which is using for faster processing for large datasets.

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. This is the first step towards student learning experiences using social media.

In future, we can use naïve Bayesian classification algorithm or decision tree algorithm to give accuracy of result. Other possible future work could analyses 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 extend to students in other majors and other institutions.

ACKNOWLEDGEMENT

We are very much thankful to our guide Prof. Prabhavathy .P who is currently working as Assistant Professor at VIT University, SITE department for timely and exclusive access to gain information on data mining and data processing issues..

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. E. Goffman, the Presentation of Self in Everyday Life. Lightning Source Inc, 1959.
4. E. Pearson, "All the World Wide Web's a Stage: The performance of identity in online social networks," First Monday, vol.14, no. 3, pp. 1–7, 2009.
5. D. Gaffney, "#iranElection: Quantifying Online Activism," in WebSci10: Extending the Frontier of Society On-Line, Raleigh, NC, 2010.
6. S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S.Lawson, "'I can't get no sleep': Discussing #insomnia on Twitter,"in Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, 2012, pp. 1501–1510.

7. M. J. Culnan, P. J. McHugh, and J. I. Zubillaga, "How large US companies can use Twitter and other social media to gain businessvalue," MIS Quarterly Executive, vol. 9, no. 4, pp. 243–259,2010.

8. 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," International Journal of Sport Communication, vol. 3, no. 4, pp. 454–471, 2010.
9. D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 241–249.

10. [10] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, pp. 1–12, 2009.
11. [11] K. Nishida, R. Banno, K. Fujimura, and T. Hoshide, "Tweet classification by data compression," in Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversity on the social web, New York, NY, USA, 2011, pp. 29–34.
12. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2010, pp. 841–842.
13. Robust Method of Sparse Feature Selection for Multi-Label Classification with Naive Bayes Dymitr Ruta Etisalat, British Telecom Innovation Centre Khalifa University, Fatima F302, PO Box 127788 Abu Dhabi, UAE Email: dymitr.ruta@kustar.ac.ae
14. S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson, "'I can't get no sleep': Discussing #insomnia on Twitter," in Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, 2012, pp. 1501–151.
15. Mining Social Media Data for Understanding Students' Learning Experiences XinChen, Mihaela Vorvoreanu, and Krishna Madhavan in ieee 2014.

