Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

Capturing Social Media Data for Understanding Students' Learning Experiences

Project Guide: Mrs. Minal Zope

Radheya Kale, Krunal Chavan, Nikhil Bhosale, Shantanu Bhasme

*Computer Engineering, AISSM's IOIT, radheyakale@gmail.com**
Computer Engineering, AISSM's IOIT, krnlchavan@gmail.com
Computer Engineering, AISSM's IOIT, bhosalenikhil93@gmail.com
Computer Engineering, AISSM's IOIT, bhasme24@gmail.com

Abstract- *Students' informal chatting done on social networking websites is informative. It sheds light into their educational experiences. Analysis of such data is a complex task. Human interpretation becomes mandatory for this task. It is the need to the day to automate this process. In this paper, we used certain data mining techniques to achieve our goal. We used engineering students' Twitter posts to understand the issues and problems faced by them in their educational experiences. When we took a survey, we found that engineering students face problems such as heavy study load, disinterest in academics and sleep deprivation. Based on this, we implemented a multi-label classification algorithm to classify tweets reflecting students' problems. This work showcases a methodology that shows how social media data can provide great knowledge regarding students' experiences.*

Keywords- *Naïve Bayes algorithm, Social Networking, Education, Computers, Web text analysis, Data mining.*

I. INTRODUCTION

SOCIAL networking websites, in which students can share their experiences, can relief their pain and such websites are twitter, facebook, etc. In our day to day life as the no. of social networking sites increases the students can share their experiences or we say the way they think can share in a casual or informal manner[1]. Student's interaction on digital environment can highly informative for administration and staff of universities. This will be very useful if student's can understand this the universities can make different strategy for risk students, they can improve the education quality and with this we can raise the student's recruitment, retention and success. A huge quantity of social data from social networking websites helps to figure out student's experiences, it also increases problems in educational purposes in the sense of social media data. If we consider the huge data volumes, the diversity of Internet slang, the unpredictability of location and timing of students posting on the web, as well as the complexity of students' experiences; pure manual analysis cannot deal with the exponentially growing scale of data. Pure pre programmed algorithms usually cannot represent in-depth sense within the data .In today's world many educational researchers uses different techniques like interviews, surveys, groups and classroom activities which helps to collect data which is mainly related to student's learning experiences. Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in[4].

These methods consumes so much time, so it cannot be recapitulate or duplicated with high frequency. The scale of such studies is generally finite. In student's occasion, about their experiences, students need to mirror on what they were thinking and doing sometime in the past, which may have become reduced over a time. The appearing fields of learning analytics and educational data mining (EDM) have concentrated on survey in which structured data obtained from course management systems (CMS), classroom technology usage, or manage online learning environments to inform informative decision-making. However, to the best of our knowledge, there

is no testing found to directly mine and study student display content from uncontrolled scope on the social web with the simple target of understanding students' learning experiences. The research objective of this study are 1) to exhibit a progress of social communication data sense-making for educational purposes, combine both qualitative study and large scale data mining ability as illustrated in Fig. 1; and 2) to investigate engineering students' casual conversations on Twitter, in series to recognize issues and difficulties students encounter in their learning experiences. We focus on engineering student's they display the posts on Twitter about difficulties in their educational experiences mainly because: 1. Engineering schools and sections have long been fighting with student recruitment and confinement matter. Engineering graduates constitute a remarkable part of the nation's future staff and have a direct collision on the country financial progress and world competency. 2. Based on understanding of matters and issues in student's life, policymakers and instructor can make more knowledgeable decisions on proper involvement and services that can useful for students overcome barriers in learning. 3. Twitter is a very popular social media site. Its content is mostly universal and very short (no more than 140 characters per tweet). Twitter provides free APIs that can be used to stream data. Therefore, we chose to initialize from analyzing students' posts on Twitter. In this paper, we went through an investigate action to discover the applicable data and Twitter hashtags (a Twitter hashtag is a word starting with a # sign, used to emphasize a topic).

II. PREVIOUS WORK

The main foundation of our paper lies in Sir Erving Goffman's theory of "The presentation of self in everyday life". It signifies the relationship between performance and life.

Twitter data has been used at several instances; may it be Iran elections or Indian Lok Sabha elections. The advantage lies in its API from which data can be easily extracted unlike other social networking websites where privacy policies are barriers in data extraction.

To be more specific about the mentioned context, we present an example.

Gaffney analyzes tweets with hashtag #iranElection using histograms, user networks, and frequencies of top keywords to compute online activism.

The benefits of this approach lies not only in politics but also in business analytics, sports, health sector and countless others. The methodologies for analysis comprise qualitative content analysis, linguistic analysis, network analysis, word clouds and histograms. Our classification model is based on inductive content analysis. We then applied this model on a fresh new dataset (test set). Therefore, we substitute human efforts with large scale automated data analysis.

III. PROPOSED METHODOLOGY

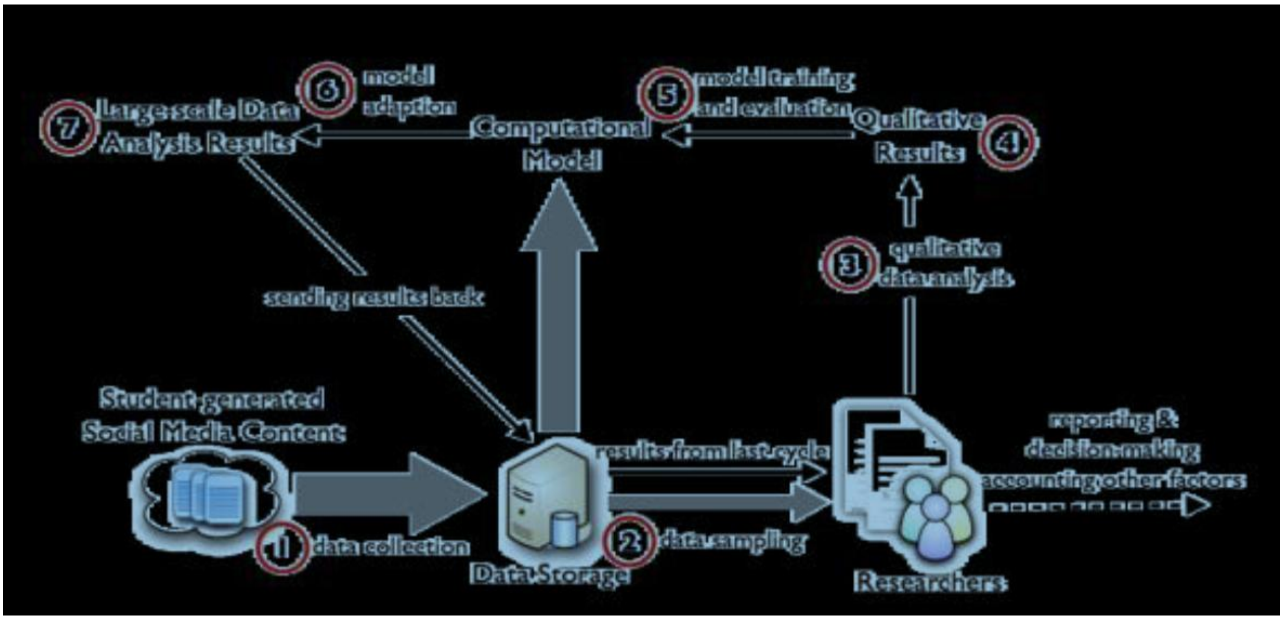

Fig 1. Architectural diagram

Naïve Bayes, Decision Tree, Logistic Regression, Maximum Entropy, Boosting, Support Vector Machine are the popular classification algorithms. There are two possible approaches based on the number of classes: Binary classification and multi-class classification. Binary classification includes only two classes whereas multi-class classification includes more than two classes.

Both binary classification and multi-class classification are single-label classification systems. Single label classification means each data point can only fall into one class where all classes are mutually exclusive. Multi-label classification, however, allows each data point to fall into several classes at the same time.

IV. INDUCTIVE CONTENT ANALYSIS

Usually, conversations on Twitter can turn out to be quite informal and might include slangs. So it requires human interpretation in order to understand the meaning of those chats.

If we intend to automate the process of large scale analysis, we need to perform deep qualitative analysis. Latent Dirichlet Algorithm (LDA) is the solution to this query. LDA is a topic modeling algorithm that can be used on a large scale for detecting general topics. But in our case, it produced word groups which were rather useless and a few even overlapped. So it became necessary to develop categories.

Top keywords found out in the analysis were sleep deprivation, heavy study load, anxiety, negative sentiments. Test dataset was analyzed post training phase.

While analyzing test data for the purpose of categorizing, a lot many tweets were found out to belong to multiple categories. That was the major reason for choosing Naïve Bayes multi-label classifier.

V. NAÏVE BAYES MULTI-LABEL CLASSIFIER

There was a major reason for choosing Naïve Bayes classification algorithm over other classification algorithms like k-nearest neighbor (KNN) algorithm and others. It was speed. Naïve Bayes algorithm was found out to be faster than KNN algorithm for large-scale data.

Naïve Bayes vs. Support Vector Machine (SVM):

Naïve Bayes was found out to be more accurate than SVM for multi-label classification.

Another algorithm with which Naïve Bayes algorithm completes is Max Margin Multi-label classifier (M3L). It was found out to be better than SVM, but Naïve Bayes was much better than M3L too.

Formula for Naïve Bayes theorem: $P(A|B) = \ P(B|A)P(A)\ P(B)$

VI. CONCLUSION

The presented study is advantageous to university administrators and/or policymakers in easing students' learning process via educational data mining. It puts forth a workflow for analyzing data from social media websites for educational purpose that overcomes the major precincts of firstly manual qualitative analysis and secondly large scale computational analysis of user-generated textual context. Educational administrators, practitioners and other relevant decision makers can gain further understanding of engineering students' college experiences.

We've put forth feasible guidelines for future work for researchers who might be fascinated in this area. Our vision regarding this research lies in all sectors. It should be a priority to protect students' privacy when attempting to provide them with high-quality education and services.

VI. FUTURE SCOPE

The approach proposed in this paper can be utilized by business analytics in taking automated survey and review of products along with necessities of potential/existing customers.

It can be used even by celebrities to overcome their shortcomings that they feel would be necessary to retain/increase their fan following.

REFERENCES

[1] Xin Chen, Student Member, IEEE, Mihaela Vorboreanu and Krishna Madhavan, "Mining social media Data for Understanding student's Learning Experiences."

[2] Dr. B. Srinivasan and P. Mekala, "Mining Social Networking Data for Classification Using Reptree"

[3] S.Cetintas, L.Si, H.Aagard, K.Bowen and M.Cordova-Sanchez, "Micro blogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging-Supported Classroom"

[4] RYAN S.J.D. BAKER,Worcester Polytechnic Institute Worcester, MA USA, KALINA YACEF School of Information Technologies University of Sydney Sydney, NSW Australia, " The State of Educational Data Mining in 2009: A Review and Future Visions "

[5] S. Shrivastav, "Data mining for Hypertext: a tutorial survey"