Identification of Student's Behavior in Higher Education from Social Media by using Opinion based Memetic Classifier

Pooja R. Takle
M.TECH Student at
Department of CST
SNDT Women's University,Usha Mittal Institute of Technology,Santacruz,Mumbai
Email ID:-pooja.takle@gmail.com

Prof. Narendra Gawai
Assistant Proffessor at
Department of CST
SNDT Women's University,Usha Mittal Institute of Technology,Santacruz,Mumbai
Email ID:- ngawai@rediffmail.com

***Abstract*:**-Social media sites such as Twitter, Facebook, You-tube are very popular social sites in higher educational student's like engineering, medical, pharmacy, trainees and other than student's also. These social media sites provides a great platform or venues for student's to share their views, emotions, stress, opinions, feelings about the learning process. Our aim is to extracting this data from social media sites for identifying the student's behavior and their opinions. The technique provides a benefit for institute or an organization to understand student's behavior. It saves our lots of time to understand the student's views. In this paper our main focus is on higher educational students for understanding student's behaviour from social media. For this we can use opinion based memetic classifier technique to integrate large scale data mining techniques and qualitative analysis to provide a better classification result of students behavior with the focus on sentimental analysis.

Keywords:-*social media, qualitative analysis, data mining, opinion, memetic classifier etc.*

*****

I. INTRODUCTION

Social networking is for everyone and it's now such a massive part of all our lives. Now a day's all over the world social media sites are very much popular such as twitter, facebook, youtube etc. It provides a great platform for students to express their views, stress, emotions, opinions, issues, joy, struggle and feelings. Student's discuss and share their everyday encounters in formal and informal manner on different social media sites. Student's tweets or comments provide large amount of implicit knowledge and a whole new perspective for educational and institutional researchers, users and practitioners to understand student's behavior outside the controlled classroom environment.

This understanding can be useful for taking the decision at institutional level by considering student's point of view for student's recruitment, retention and success. The social media data provides lots of opportunities to understand student's behavior, but also there are some methodological difficulties in making sense of social media data for educational purposes. There are number of methods that are used by educational researchers such as surveys, interviews, focus groups, and classroom activities to collect data related to student's behavior. These methods are usually very time-consuming, thus cannot be duplicated or repeated with high frequency [1]. The emerging fields of learning analytics and educational data mining (EDM) have focused on analyzing structured data obtained from course management systems (CMS), classroom technology usage, or controlled online learning environments to inform educational decision-making [2],[5], [6], [7]. However, as per our knowledge, this is the first research found to directly mine and analyze student posted content from uncontrolled spaces on the social web with the clear goal of understanding students' learning behavior.

The main research goals of this technique are 1) to develop a work flow of social media data for higher educational purposes, integrating both large scale data mining techniques and qualitative analysis in Fig.1 to extract higher educational student's formal and informal conversations on Twitter, facebook in short social media sites in order to understand issues, problems, views, opinions, stress, emotions student's encounter in their learning behavior/experiences. We chose to focus on higher educational students' posts on Twitter about problems in their educational behavior mainly because:

1.Higher educational schools and departments have long been struggling with student recruitment and retention issues [8]. Higher educational graduates constitute a significant part of the nation's future workforce and have a direct impact on the nation's economic growth and global competency [9].

2. Based on understanding of issues and problems in students' life, policymakers and educators can make more informed decisions on proper interventions and services that can help students overcome barriers in learning [1].

3. Twitter is a popular social media site. Its content is mostly public and very concise (no more than140 characters per tweet). Twitter provides free APIs that can be used to stream data. Therefore, we chose to start from analyzing student's posts on Twitter [1].

The remaining part of this paper is organized as follows. In

Section 2, we describe the literature survey on dentification of students learning behavior by extracting the social media data. In Section 3 describe the architecture of the system and in Section 4-opinion based memetic classifier.In section 5 describe a privacy preservation techniques after that section 6 and 7 discusses the benefits and drawbacks of the system.In section 8 and 9 concludes this study and discusses the possible future work.

II. LITERATURE REVIEW

There are different techniques have been developed for extracting the datasets through social media such as Radian6 tool(http://www.salesforce.com)[1] and_FoursquareAPI which help for extracting the data, and that data is available in the .csv,.xls file. Twitter offers a set of APIs for retrieving the data about its users and their communication [3].

The theoretical foundation for the value of informal data on the web can be drawn from Goffman's theory of social performance [10]. When student's wants to post any content on social media sites, they usually post what they think and feel at that moment. In this sense, the data collected from online conversations may be more authentic and unfiltered than responses to formal research prompts. These conversations act as a outlook for student's behavior. Many studies show that social media users may purposefully manage their online identity to "look better" than in real life [12], [13]. Other studies show that there is a lack of awareness about managing online identity among college student's [14], and that young people usually regard social media as their personal space to hang out with peers outside the sight of parents and teachers [15]. Student's online conversations reveal aspects of their behaviors that are not easily seen in formal classroom settings.Popular classification algorithms include Naive Bayes,Decision Tree, Logistic Regression, Maximum Entropy, Boosting, Support Vector Machine (SVM), C4.5, J48 etc. are use to classify the social media data. Sentiment analysis is another very popular three-class classification on positive, negative, or neutral emotions/opinions[18]. Sentiment analysis is very useful for mining customer opinions on products or companies through their reviews or online posts. It finds wide adoption in marketing and customer relationship management (CRM).

1. NAIVE BAYES MULTI-LABEL CLASSIFIER

The Naive Bayes Multi-label Classifier is totally a probability based. The Naive Bayes is often used as a baseline in text classification because it is easy to implement. This classifier is very effective on the dataset for classify the tweets based on the categories. It uses the text preprocessing technique to avoid the repetation of the text. The Naive Bayes Multi-label Classifier is one popular way to implement multi-label classifier is to transform the multi-label classification problem into multiple single-label classification problems. One simple transformation method is called one-versus-all or binary relevance [19].

The basic concept of this binary relevance is to assume independence among categories and train a binary classifier for each category. After that all binary classifiers can be transformed to multi-label classifier using the one versus all heuristics. The basic procedure is, in the training document collection there are total number of n words and in this case each tweet or comment is a document.

w={w1,w2,w3,......wn } and all the tweets are related to different categories c={c1,c2,c3..cL} with total number of L categories c.

2. CRISP-DM METHODOLOGY

Data Mining is the analysis step of the Knowledge Discovery in Databases process (KDD). Data Mining is the process using analysis techniques of computer automated to extracting the knowledge from data . It is the computational process of discovering patterns in large datasets. The main goal of Data Mining (DM) process is to extract the knowledge from datasets and then transform or convert it into an understandable format or structure for next use.CRISP-DM Methodology is a Cross Industry Standard Process for Data Mining. This CRISP-DM methodology takes different stages;

Stage 1: Business Understanding

The main purpose and goal of this first stage of CRISP-DM methodology is to define the objectives and reasons of Knowledge Discovery Databases(KDD) process.

Stage 2: Data Understanding

This is the second stage of CRISP-DM methodology. It includes with the collecting and transforming the data into a format that can be used by the selected data mining tools, data description, data exploration and verification of data quality [17].

Stage 3: Data Preparation

This data preparation stage is necessary to prepare the data.

This stage is conducted through the tasks of data selection, data cleansing, data construction, integration and formatting data, dataset, data preparation stage and data description will be used in a data modelling.

Stage 4: Modelling

Selection of techniques is the initial step of modeling stage to build a data mining model.

There are different techniques for different problems such as classification, clustering, prediction, regression etc. This techniques are depends on data quality, time, fact.

Stage 5: Evaluation of Results

The CRISP-DM Methodology refers to the evaluation of results in the context of the extent to which the model meets the business objectives and the results generated using data mining methods in order to select the model that will be applied [17].

If the results are not satisfactory then it is necessary to return back to the previous stage of modelling.

Stage 6: Implementation Strategy

The last stage of CRISP-DM Methodology is implementation strategy of the results of data mining analysis in order to improve business and its totally based on the evaluation results.

The main purpose and goal is to create general procedures for creating a relevant model.

III. ARCHITECTURE

The main goal or a purpose of this architectural process is to extracting the students behavior from social media sites, analyze them and taking a final decision related to students behavior - exactly what they wants?

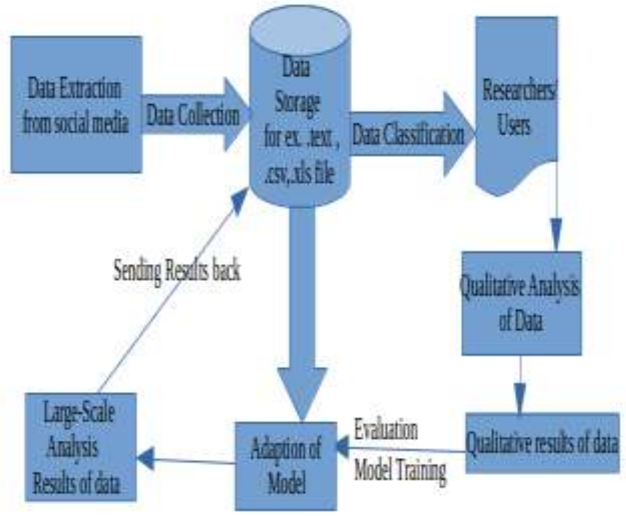
Data Extraction from social media
Data Collection
Data Storage for ex. .text , .csv,.xls file
Data Classification
Researchers/ Users
Qualitative Analysis of Data
Qualitative results of data
Evaluation
Model Training
Adaption of Model
Large-Scale Analysis Results of data
Sending Results back

Figure 1:- Architecture of extracting students behavior from social media data and analyze them.

This work-flow we can developed for making sense of social media data integrates qualitative analysis and data mining algorithm. The work-flow can be an iterative cycle [1].

where,
blue arrows:- represents data volumes
wider blue arrows:- more data volumes
simple blue arrows:- computation, data analysis and results flow

As per the architectural process, to takes a final decision having the different types of stages;

Stage 1: Data Extraction

In this stage there are different ways to collect the tweets from social media sites such as FoursquareAPI, Radian6 Tool, Public Twitter API, www.salesforce.com etc. The datasets of tweets generated using the geocode(longitude and latitude) of our area and this dataset is in the form of .csv, .xls, .txt file i.e a data storage.

Stage 2: Data Classification

In this stage an inductive content analysis on samples or procedures of the #learning problems dataset. This type of procedure is called as data sampling.

Stage 3: Qualitative Analysis of data

In this stage of qualitative data analysis it identifies the categories.

Stage 4: Qualitative Results of data

As per the stage 3 of qualitative data analysis it identifies the categories and then in stage 4 it returns the keywords as per the categories.

Stage 5: Evaluation and Model Training

As per the categories we can implement a opinion based memetic classifier(explained in section 4) to analyze and classify the tweets.

Stage 6: Adaption of Model

In this stage the classification algorithm to train a detector that could assist detection of higher educational students problems.

Stage 7: Large-Scale Analysis Results of data

this is the last stage of the process, it returns the total final results in the classification form with the sentimental analysis.

As per this process makes two major contributions, 1. it proposes a work-flow to integrate and bridge a large scale data mining techniques and qualitative research methodology.

2. it provides deep insights into higher educational students educational experiences as reflected in informal, uncontrolled environments.

The higher educational students have lots of issues and problems such as lack of social engagement, sleep problems, study-life-balance, diversity issues and others. Our proposed technique we will be able to solve such types of problems.

IV. OPINION BASED MEMETIC CLASSIFIER

The basic memetic algorithm is an extension of genetic algorithm by a local search. The memetic algorithm is a population based approach. The population based global technique and a local search made by each of the individuals. In a memetic algorithm the population is initialized at random or using a heauristic. The genetic and a memetic algorithms are mostly used for optimization. The memetic algorithm returns a proper optimization result than genetic algorithm. After the complition of a optimization procedure of memetic algprithm applying or appending a classification technique to it. We have develope a new algorithm/classifier and we call it as an Opinion based Memetic Classifier. This opinion based memetic classifier can returns a results with fully optimized classification results.

ALGORITHM:-

```
1. Encode solution space
2. a) set pop_size, max_gen, gen=0
   b) set cross_rate, mutate_rate
3. Initialize population
4. while( gen< gensize )
        apply generic genetic algorithm
        apply local search
   end while
5. Apply final local search to best chromosome
6. Apply classification technique
```

as per this algorithm we can apply a hill climbing or any local search to it. And on the step 6 of this algorithm, we can apply any classification technique to it like C4.5, J48. These are the simple classification techniques to generate the decision tree. This decision tree of binary value and after that it returns a results into the form of classification chart and pie chart for sentimental analysis.

V. PRIVACY PRESERVATION ON DATA

Privacy preservation is also a important part of this. We can apply the different techniques on data/dataset of the students so no one can hack it or modify it.

There are different types of privacy preservation techniques;

- **Shuffling technique:**

Data shuffling technique is similar to a substitution.In the database the data is randomly shuffled within the column.The shuffling technique is use for hiding the original data with random data for privacy preservation.

- **Transferring random data from datasets:**

In this technique the data is directly transfer with random data from the datasets or database.

- **Encryption:**

Encryption is the most effective way to achieve data security.Encryption is the process of encode the original data into unreadable format by using cryptographic techniques like substitution techniques,transposition techniques for exp. Caesar cipher, monoalphabetic cipher, playfair cipher, hill cipher etc.

VI. ADVANTAGES

1. Identifying students learning behavior from social media system provides a privacy preservation on data.

2. This system also provides the classification results with the sentimental analysis of the behaviors of students data.

3. This system implementing a new algorithm i.e Opinion based Memetic Classifier . In this algorithm the basic memetic algorithm is used which is only used for optimization now a days but in this system it uses as a classifier so will get a better optimized results.

VII. LIMITATIONS

1. This System is only used for text data or structured data not for unstructured data like videos, images etc.

VIII. CONCLUSION

The above discussed classification techniques for understanding students behavior like Naive Bayes Multi-label Classifier and CRISP-DM Methodology are good. This survey paper on identification of students learning behavior from social media can be helpful for finding the drawbacks of existing classification algorithm. The new Opinion based Memetic Classifier algorithm would definitely help in developing a new system that combines all the advantages and overcomes the drawbacks of existing systems.

IX. FUTURE WORK

This "Identification of Studen's Behavior in Higher Education from Social Media by using Opinion based Memetic Classifier" system will use videos,images,notations or smilies etc. in the future for better results.

X. REFERENCES

[1] Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, Krishna Madhavan,"Mining Social Media Data for Understanding Students' Learning xperiences," IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO.3,JULY-SEPTEMBER2014.

[2] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representation and Communication: Challenges in Interpreting Large Social Media Datasets," Proc. Conf. Computer Supported Cooperative Work,pp. 357-362, 2013.

[3] OLAPing Social Media: The case of Twitter by Nafees Ur Rehman, Andreas Weiler, Marc H. Scholl University of Konstanz, Germany Email: {nafees.rehman, andreas.weiler, marc.scholl @uni-konstanz.de}, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[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] R. Ferguson, “The State of Learning Analytics in 2012: A Review and Future Challenges,” Technical Report KMI-2012-01, Knowledge Media Inst. 2012.

[6] 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, 2009.

[7] 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. Learning Technologies,vol. 4, no. 4, pp. 292-300, Oct.- Dec. 2011.

[8] C. Moller-Wong and A. Eide, “An Engineering Student Retention Study,” J. Eng. Education, vol. 86, no. 1, pp. 7-15, 1997.

[9] National Academy of Eng., The Engineer of 2020: Visions of Engineering in the New Century. National Academies Press, 2004.

[10] E. Goffman, The Presentation of Self in Everyday Life. Lightning Source Inc., 1959.

[11] 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.

[12] J.M. DiMicco and D.R. Millen, “Identity Management: Multiple Presentations of Self in Facebook,” Proc. the Int’l ACM Conf. Supporting Group Work, pp. 383-386, 2007.

[13] M. Vorvoreanu and Q. Clark, “Managing Identity Across Social Networks,” Proc. Poster Session at the ACM Conf. Computer Supported Cooperative Work, 2010.

[14] M. Vorvoreanu, Q.M. Clark, and G.A. Boisvenue, “Online Identity Management Literacy for Engineering and Technology Students,” J. Online Eng. Education, vol. 3, article 1, 2012.

[15] M. Ito, H. Horst, M. Bittanti, D. boyd, B. Herr- Stephenson, P.G. Lange, S. Baumer, R. Cody, D. Mahendran, K. Martinez, D. Perkel, C. Sims, and L. Tripp, Living and Learning with New Media: Summary of Findings from the Digital Youth Project. The John D. and Catherine T. MacAuthur Foundation, Nov. 2008.

[16] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up?: Sentiment Classification Using Machine Learning Techniques,” Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing, vol. 10, pp. 79-86, 2002.

[17] Grljevic Olivera , Bosnjak Zita , Bosnjak Sasa ,”Students’ Behavior on Social Media Sites – A Data Mining Approach” University of Novi Sad, Faculty of Economics Subotica/Business Information Systems and Quantitative Methods, Subotica, Serbia oliverag@ef.uns.ac.rs,bzita@ef.uns.ac.rs, bsale@ef.uns.ac.rs, SISY2013•IEEE 11th International Symposium on Intelligent Systems and Informatics • September 26-28, 2013, Subotica, Serbia

[18] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up?: Sentiment ClassificationUsingMachineLearning Techniques,” Proc.ACL-02Conf.Empirical Methods in NaturalLanguageProcessing,vol.10,pp.79- 86,2002.

[19] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining Multi-Label Data,” Data Mining and Knowledge Discovery Handbook, pp. 667-685, Springer, 2010.

AUTHORS PROFILE

Prof. **Narendra Gawai** received his B.E. in Computer Engineering from Government College of Engineering, Amravati, Amravati University, M.E. from VJTI, Mumbai University, Maharashtra, India. He is currently working as Assistant Professor at UMIT, SNDT Women’s University, Mumbai. He has 15 years of teaching experience. He has guided several undergraduate projects. His area of interests are Systems Security, Digital Forensics, Big Data Analytics, Data Warehousing and Data Mining.

Pooja Takle received her B. E degree in Computer Science and Engineering, from NDMVP's K.B.G.T College of Engineering, Nashik of Pune University. She is currently pursuing her M. Tech(2nd Year) in Computer Science and Technology from UMIT, SNDT Women's University, Mumbai. Her area of interests are Data Mining, Networking.