ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Interpreting Students Behavior using Opinion Mining

Pooja R. Takle,Prof. Narendra Gawai

M.Tech Student, Dept. of CST, UMIT, SNDT Women's University, Santa Cruz, Mumbai, India

Assistant Professor, Dept. of CST, UMIT, SNDT Women's University, Santa Cruz, Mumbai, India

ABSTRACT:-There is a increase uses of social media sites like twitter, Facebook, you-tube,google+ by students of higher education. All these sites allow users to keep in touch with friends, relatives, neighbors, colleagues etc. On the social networking sites the data is available in structured and unstructured manner. This type of data provides valuable knowledge related to higher educational students actions or behavior. To analyze such type of data and getting the information related to students is usefulto improve or enhance the higher education system. This paper describes analysis and classification of structured databy using classification algorithm. In the existing system Naïve Bayes Multi-label Classifier is used. This classifier gives a good results but its very time consuming. So, to overcome this limitation we proposed a technique called as a "Memetic Classifier" based on genetic algorithm. We have done comparative study of classification techniques such as Iterative dichotomiser3 (ID3),Nave Bayes Multi-label Classifier and memetic Classifier using common dataset. By using Memetic Classifier, we are able to identify the student's behavior properly.

KEYWORDS:- Iterative Dichotomiser3, Naïve Bayes Multi-label Classifier, Genetic algorithm, Memetic Classifier, Opinion mining.

I. INTRODUCTION

Social media or social networking sites plays a very important role in our daily life. In today's fastest life we mostly use websites for entertainment, business, shopping, social networking sites like twitter, Facebook, YouTube, google+ etc. and education and many more. All these type of sites are very much popular in youngsters and most of them are higher educational students. It provides a great venues for youngsters to share their opinions, feelings emotions, views, stress, issues, struggle, joy about the learning process and other. The higher educational students share and discuss their everyday daily routine in formal and informal manner.

The social media data provides lots of opportunities to understand student's actions or behavior. The educational researchers used various manual methods like focus groups, personal interviews classroom activities, their behavior, actions, survey's etc. to collect student's related data. All these types of manual methods are very time consuming. By considering the disadvantage of manual testing,the new system is developed. This understanding is very useful for taking the decision at university or organizational level by considering student's point of view for student's success. This newly developed system Interpreting Students Behavior using Opinion Mining overcomes all these drawbacks of existing system. In existing system Naïve Bayes Multi-label classifier is used. This Naïve Bayes Multi-label classifier gives good results with all probabilistic values but it is very time consuming and also in the existing system the part of sentimental analysis is not present. In the newly developed system we implemented a new classifier i.e Memetic Classifier. This Memetic Classifier gives the better results than the others in a very short period of time. The rising fields of Educational data mining and learning analytics have focused on analyzing structured data obtain from classroom technique usage, Course Management System (CMS), Online Learning environments to inform decision making.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

However, as per our knowledge , this is the first research found to directly mine and analyze students posted comment or text or content on social wed or on social networking sites with the clear goal of recognition or identification of higher educational students learning actions or behavior.
Find out and classify higher educational student's issues and problems in their learning. Find out the data track or path of higher educational students with the good or bad actions or behavior.Mining this social media data like their actions or behaviors will result to classify the group of higher educational students according to their actions and recognize their issues and problems to be solved to improve the higher educational quality. Sometimes in the higher education some students are not from convent so they gets shy or afraid of clearing their issues and problems in the classroom. So, at this time many students prefers the social media like twitter, Facebook, You-tube etc. This social media helps them to post or share their views, emotions, opinions, stress, feelings about their learning process. The higher educational departments have been struggling with student's recruitment and retention issues. This Interpreting Students Behavior using Opinion Mining new system is very useful to higher education to improve their education quality.The remaining part of this paper is organized as follows. In section 2 literature review is explained with ID3 and Naïve Bayes Multi-Label Classifier. In section 3 overview of proposed work with newly implemented algorithm i.e Memetic Classifier and systems architecture and in section 4 Results and analysis are discussed.

II. LITERATURE REVIEW

Social media plays a very important role in our day to day life. Everyone or from students to teachers, friends, youngsters, VIP authorities all are registered on social media sites. These social networking sites are user friendly so everyone can share or discuss their views, emotions, actions, issues, stress, feelings, opinions etc.
There are different techniques for extracting the data of higher educational students through social media such as Radian 6 Tool-www.salesforce.com, FourSquareAPI, PublicTwitter API etc.

ID3(Iterative Dichotomiser 3):-
Iterative Dichotomiser 3 decision tree algorithm is based on the Concept Learning System(CLS) algorithm. This ID3 developed at University of Sydney by Ross Quinlan. This algorithm is used to generate a decision tree from a large dataset and typically used in Natural Language Processing(NLP) and in Machine Learning domains. ID3 uses a greedy approach to find out a better attributes to split the dataset on each node or iteration to form a decision tree. ID3 uses entropy and information gain to construct or build a decision tree.

Entropy:-
Entropy is the average value or expected value of the dataset or information. It is one kind of measurement procedure in information theory. Entropy is used in ID3 algorithm to calculate the homogeneity of a sample. When,
if entropy=0 i.e sample is completely homogeneous
if entropy= 1 i.e sample is an equally devided
Entropy(S) $=-a\log_2 a - b\log_2 b$

Where, S = Sample Records
a = data of set 1 or portion of set 1
b = data of set 2 or portion of set 2

Information gain:-

Information gain uses entropy to calculate effective change in entropy after making a decision based on the value of an attribute Q. It's ideal to base decisions on the attribute that provides the largest change in entropy and this attribute with the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

highest gain, for decision trees. The information gain is the measure of difference in entropy from before and after, the set S is split on an attribute Q. Information gain (S, Q) of example set S on attribute Q,

$$I_{Gain}(S,Q) = \text{Entropy}(S) - S((|S_p| / |S|) * \text{Entropy}(S_p)$$

Where,
S I each value v of all possible values of attribute Q
S_p = subset of S for attribute Q has value p
$|S_p|$= number of elements in S_p
|S|= number of elements in S

Gain quantifies the entropy improvement by splitting over higher attribute.

Naive Bayes Multi-label Classifier:-

Naïve Bayes Multi-label classifier is a probabilistic classifier. It's totally a probability based classifier. This classifier is easy to implement and it's often used as a baseline in text classification. It is good to classify the tweets based on categories. Text preprocessing technique is used to avoid repetition of text. Multi-label classifier is to transform multi-label classification problem into multiple single label classification problem, this is the popular way to implement it. One-versus-all or binary relevance is one simple transformation method and this basic concept is to assume independence among categories, and train a binary classifier for each category. There are total number of N words in training document and in our case each tweet or text or comment is a document.
Therefore W= $\{w_1, w_2, w_3, \ldots\ldots\ldots\ldots, W_N\}$,
Categories C= $\{C_1, C_2, \ldots\ldots\ldots\ldots, C_L\}$ in total number of L categories C.
Therefore W_{di} = $\{ W_{i1}, W_{i2}, W_{i3}, \ldots\ldots\ldots\ldots, W_{ik} \}$
Where, di = document in the testing set
k= words
W_{di} = subset of W

Therefore according to Baye's theorem,

$p(C|d_i) = p(d_i \mid C).\ p(C) / p(d_i) \propto \prod_{k=1}^{k} p(W_{ik} \mid C) . p(C)$[1]
Probability that d_i belongs to Category C

$p(C'|d_i) = p(d_i \mid C') .\ p(C') / p(d_i) \propto \prod_{k=1}^{k} p(W_{ik} \mid C') . p(C')$ [1]
Probability that d_i belongs to categories other than C

The Naïve Bayes Multi-label classification problem is transformed into the combination of several single label Naïve Bayes classifiers and this algorithm is efficient for real time systems.

III. PROPOSED SYSTEM

The proposed System "Interpreting Students Behavior using Opinion Mining" is very useful to understanding the behavior of higher educational students from social media. This system works faster than Existing system. For getting the better results we developed a new algorithm i.e Memetic Classifier. This system provides the accurate results with the opinion analysis.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Memetic Classifier:-

The memetic algorithm is a population based approach and it's an extension of genetic algorithm. There are mostly used for optimization. The memetic algorithm is extended by a local search. The genetic algorithm begins with a randomly selected population of chromosomes and the population is initialized at random or using a heuristic. The memetic algorithm refers to hybrid algorithms, a marriage between a population based global technique and a local search made by each of the individuals.
The new implemented or developed memetic classifier is an extension of a memetic algorithm by by a classification technique. After the completion of optimization procedure of memetic algorithm applying or appending a classification technique to it. This memetic classifier returns a fully optimized classification results in very less time as compare to naïve bayes multi-label classifierand ID3(Iterative Dichotomiser 3 algorithm. We can append any classification technique to optimized results of memetic algorithm like C4.5,J48, naïve Bayes Multi-label classifier, Iterative Dichotomiser 3 (ID3) algorithm. But we used naïve Bayes Multi-label classifier. Henceforth the memetic classifier is better than others.

Pseudocode:-

-Encode Solution Space with pop-size,max-gen,gen=0,cross-rate,mutate-rate
-generate an initial population of solutions of chromosomes
while(gen<gensize) for fitness evaluationby applying generic genetic algorithm
-for(i=1 to pop-ssize)
select(mate1,mate2) the parents that are to be select for reproduction according to the fitness
-if(rnd(0,1) \leqcrosss-rate)
child=crossover(mate1,mate2) for crossover to select high fitness chromosomes then few chromosomes of one parent is replace with other parent so that new offspring should be generated.
-if(rnd(0,1)\leq mutate-rate)
child=mutation(); repair child if necessary, applying for low probability after crossover two same offsprings are produced then random bits from one offspring is mutated to produce different offspring
-applying local search starting from each member of the offspring set
-replace the lowest solutions in population with new offspring
- again implementation of mutation parameter until get better optimized results
-applying naive bayes multi-label classifier to that better optimized results

IV. ARCHITECTURE

The main goal of this architectural process is to extraction of the tweets or comments or text or data from social media of higher educational students for identifying the students actions or behavior related to study, analyze them and takes a final decision with the opinion analysis related to students- what they wants exactly?

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

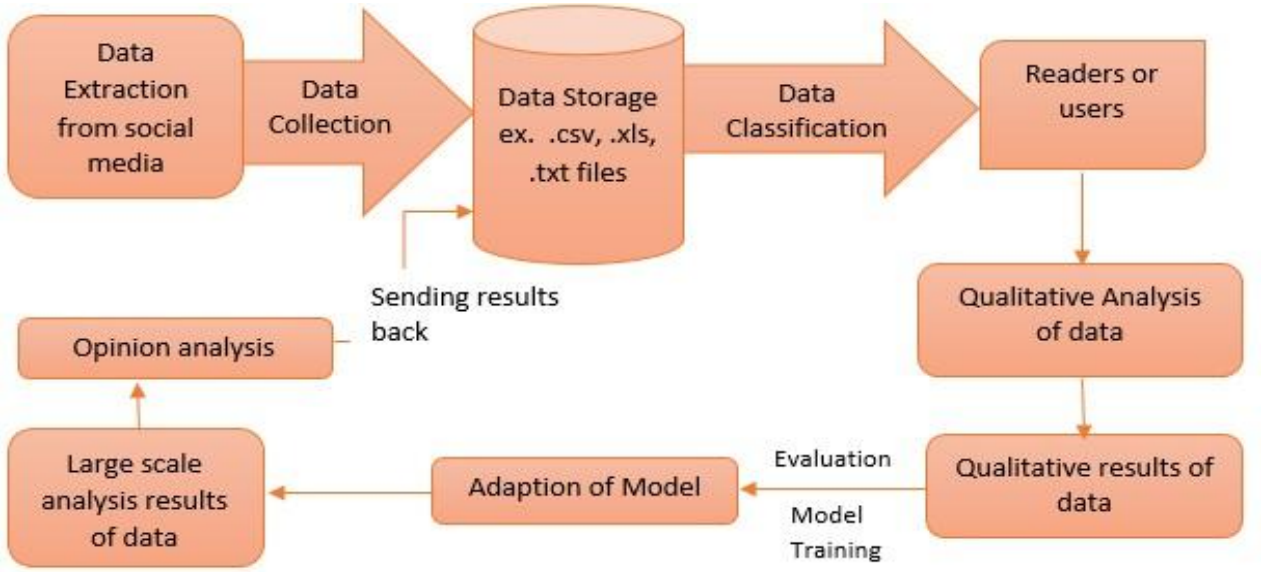


Figure:- Architecture of Interpreting Students Behavior using Opinion Mining

This architecture diagram shows the workflow of proposed system i.e "Interpreting Students Behavior using Opinion Mining ."
In this, Light red arrows:- represents data volumes
Simple light red arrows:- computation, data analysis, and results flow
Wider light red arrows:- more data volumes

This architectural process having with the different stages;
Extraction of students data:-

For identifying the students actions or behavior from social media, the extraction of students data from social sites is very important part of this process. There are different ways to collect the tweets or comments. We can use Radian 6 tool, Public TwitterAPI, FourSquare API, www.salesforce.com etc. This dataset is in the form of .csv, .xls, .txt files and i.e a data storage.

Data Classification or Data Sampling:-

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[2].
Qualitative analysis of students data:-

In this stage it identifies the categories of students data.
Qualitative results of students data:-
It returns the keywords as per the categories after the identification of categories.
Model training and evaluation:-

In this stage the main classifier or newly implemented classifier i.e memetic classifier is executed for to analyze and classify the tweets.
Adaption of model:-

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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Large scale analysis results of data:-
In this stage it returns the categorized result in the percentage form as per the categories of student data.
Opinion Analysis / Opinion Mining / Sentimental Analysis:-
In this stage it returns the overall final result in the positive and negative form. This process is also called as sentimental analysis.

The higher educational students have lots of issues and problems related to their studies and educational system such as heavy study load, lack of social engagement, negative emotion, sleep problems, diversity issues and other etc. This proposed system is able to solve such types of problems.

V. RESULTS AND ANALYSIS

1) **Comparison between Iterative Dichotomiser3, Naïve Bayes Multi-label Classifier and memetic classifier**
Input: dataset of student's data or comments
Output: analysis of ID3,NB and MA classifiers with accuracy

This result shows that the Memetic classifier is better than ID3 and NB classifiers

Graph:- Comparison between ID3,NB,MA Classifiers

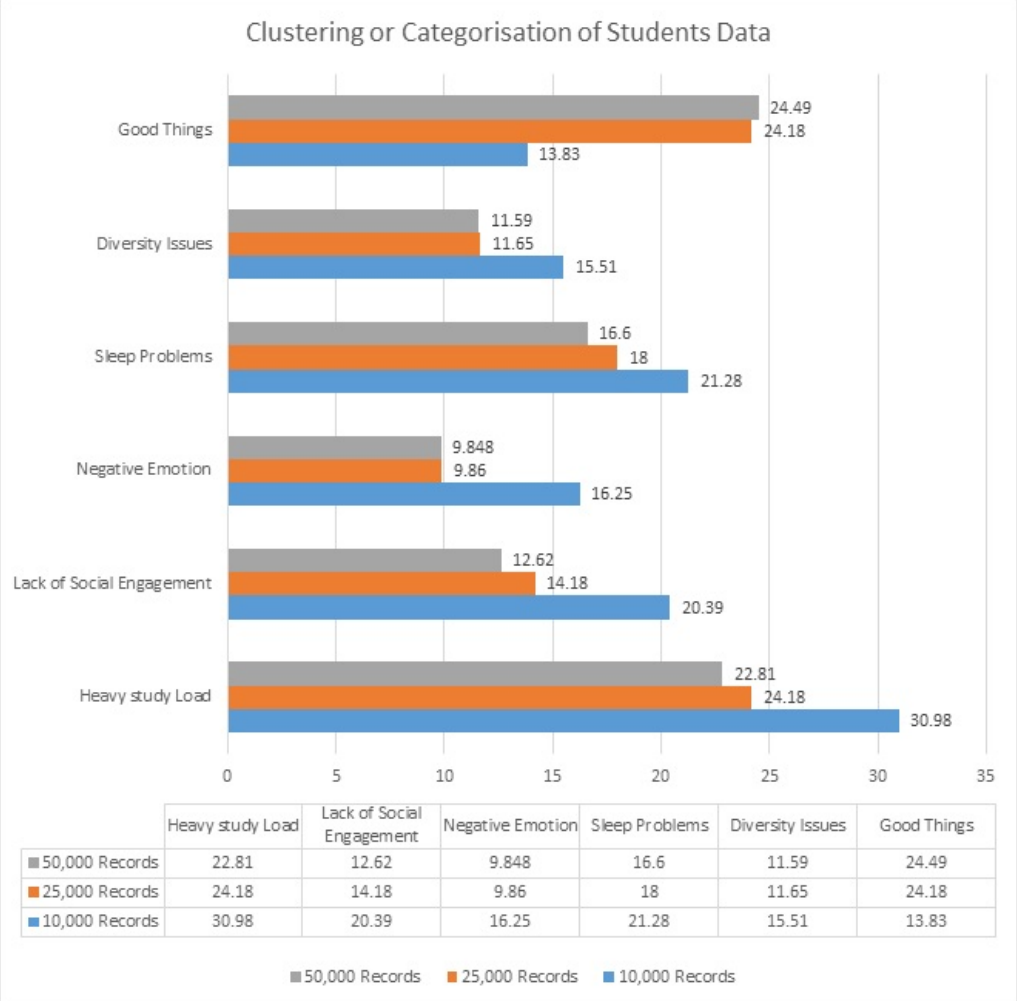
2) **Tweets in the percentage form as per the categories on each dataset.**
Input: Input Dataset file of Students data or comments
Output:- displaying categorization or clustering of students problems and issues with 6 different categories

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015



	Heavy study Load	Lack of Social Engagement	Negative Emotion	Sleep Problems	Diversity Issues	Good Things
50,000 Records	22.81	12.62	9.848	16.6	11.59	24.49
25,000 Records	24.18	14.18	9.86	18	11.65	24.18
10,000 Records	30.98	20.39	16.25	21.28	15.51	13.83

Graph:- Clustering or categorization of students data

3) Opinion analysis of database of all datasets or full data storage

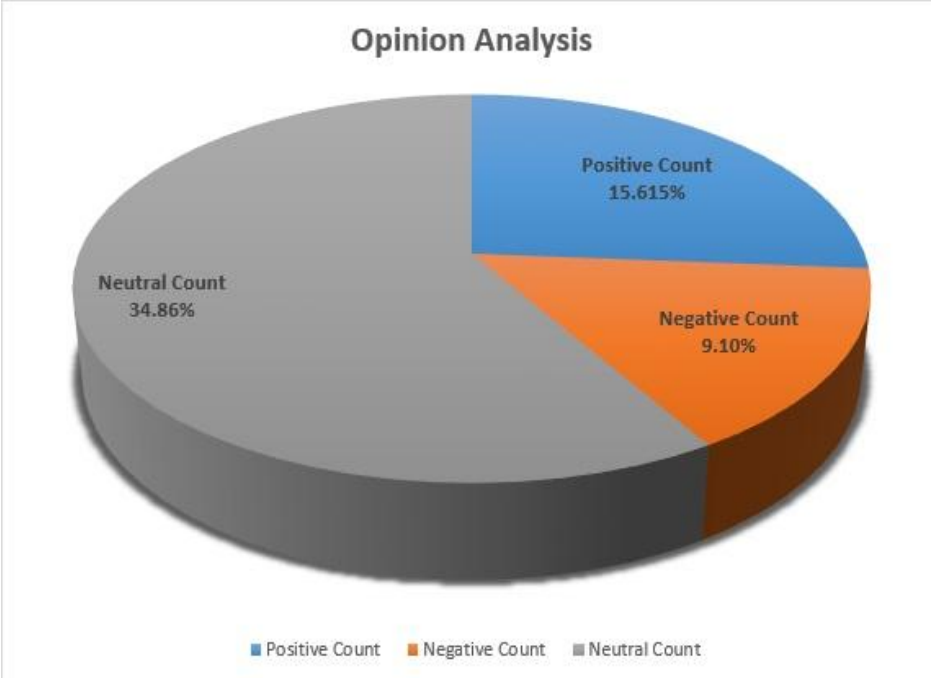
Input:- Input database of dataset files(Total Records:- 1,10,000 records) of students data or comments
Output:- displaying results with positive , negative and neutral count

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015



Graph:- Opinion Analysis of students data

VI. ADVANTAGES

1. Interpreting Students Behavior using Opinion Mining 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. The newly implemented algorithm i.e Memetic Classifier . In this classifier the basic memetic algorithm is used which is only used for optimization nowadays but in this system it used as a classifier so we get a better optimized results.

VII. LIMITATIONS

Presently we have extracted structured data from various social sites to analyze the student's behavior.

VIII. IMPLICATIONS

1Interpreting Students Behavior using Opinion Mining system is very useful in institutes, organizations and Universities to identify student's behavior in learning process.
2. This system is also very useful for industry, manufacturing companies, banking sectors, government sectors etc. in future for identification of employees actions, their behaviors, product feedback, for banking feedback and related to their services.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

IX. CONCLUSION AND FUTURE SCOPE

This paper described the newly Interpreting Students Behavior using Opinion Mining is good to use as compare to manual qualitative analysis and existing system. This system implemented with a new classifier called Memetic Classifier.In the existing system Naive Bayes Multi-label Classifier is used and this classifier is very time consuming. From this work done and survey on Interpreting Students Behavior using Opinion Mining is helpful to found the drawbacks of existing classification algorithm. The newly implemented Memetic Classifier is analyzes the very large amount of data in a short period of time. This newly developed system provides the categorization or clustering of student problems and issues and also provides the opinion analysis of database. As per the survey and work done this system very useful in institutes, organizations, universities. We can enhance in image processing like images, emoticons, videos etc. This system is also very useful for industry, manufacturing companies, banking sectors, government sectors etc. in future for identification of employees actions, their behaviors, product feedback, for banking feedback and related to their services.

REFERENCES

[1] Xin Chen, Student Member, IEEE, MihaelaVorvoreanu,KrishnaMadhavan,"Mining Social Media Data forUnderstanding Students' Learning experiences," IEEETRANSACTIONS ON LEARNING TECHNOLOGIES,VOL. 7, NO.3,JULY-SEPTEMBER2014.
[2] Students Behavior on Social Media Sites A Data Mining Approach by GrljevicOlivera*, Bosnjak Zita*, BosnjakSasa* SISY 2013 IEEE 11th International Symposium on Intelligent Systems and Informatics September 26-28, 2013, Subotica, Serbia
[3] Classification of Student's E-Learning Experiences' in Social Media via Text Mining by Ms. Priyanka Patel1, Ms. Khushali Mistry2 ,Department of CSE, PIET, Vadodara, India,IOSR Journal of Computer Engineering (IOSR-JCE),e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. VI (May – Jun. 2015), PP 81-89
[4] 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.
[5] 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 Minin.
[6] 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.
[7] R. Ferguson, "The State of Learning Analytics in 2012: A Review and Future Challenges," Technical Report KMI-2012-01, Knowledge Media Inst. 2012.
[8] 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.
[9] 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. LearningTechnologies,vol. 4, no. 4, pp. 292-300, Oct.- Dec. 2011.
[10] C. Moller-Wong and A. Eide, "An Engineering Student Retention Study," J. Eng. Education, vol. 86, no. 1, pp. 7-15, 1997.
[11] National Academy of Eng., The Engineer of 2020: Visions of Engineering in the New Century. National Academies Press, 2004.
[12] E. Goffman, The Presentation of Self in Everyday Life.Lightning Source Inc., 1959.
[13] 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.
[14] 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.
[15] M. Vorvoreanu and Q. Clark, "Managing Identity Across Social Networks," Proc. Poster Session at the ACM Conf. Computer Supported Cooperative Work, 2010.
[16] 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.
[17] 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.
[18] 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.
[19] GrljevicOlivera ,Bosnjak Zita , BosnjakSasa ,"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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

[20] 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.
[21] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-Label Data," Data Mining and Knowledge Discovery Handbook, pp. 667-685, Springer, 2010.