Opinion mining from student feedback data using supervised learning algorithms

Dhanalakshmi V., Dhivya Bino
Faculty, Department of Computing
Middle East College
Muscat, Oman

Saravanan A. M.
Faculty
Caledonian College of Engineering
Muscat, Oman

***Abstract*—This paper explores opinion mining using supervised learning algorithms to find the polarity of the student feedback based on pre-defined features of teaching and learning. The study conducted involves the application of a combination of machine learning and natural language processing techniques on student feedback data gathered from module evaluation survey results of Middle East College, Oman. In addition to providing a step by step explanation of the process of implementation of opinion mining from student comments using the open source data analytics tool Rapid Miner, this paper also presents a comparative performance study of the algorithms like SVM, Naïve Bayes, K Nearest Neighbor and Neural Network classifier. The data set extracted from the survey is subjected to data pre-processing which is then used to train the algorithms for binomial classification. The trained models are also capable of predicting the polarity of the student comments based on extracted features like examination, teaching etc. The results are compared to find the better performance with respect to various evaluation criteria for the different algorithms.**

Keywords—; Sentiment analysis; Opinion mining; Supervised learning; Machine Learning; Natural Language Processing; Rapid Miner; Text Analytics ;Learning Analytics

I. INTRODUCTION

The amount of data that any organization in the present day world is dealing with is massive. This "Big Data" is normally characterized by 3 V's i.e. volume , velocity and variety. It is often more about all these 3 dimensions coming together than considering any of them separately. Most of this data is unstructured existing in the form of emails, images, and weblogs and so on. According to a recent study, the percentage of this unstructured data amounts to around 80 of the total [1].Unfortunately, most of the decision making is done based on analysis of the structured data leaving the analytics of the 80% still at the back door.

Wassan argues in [2] that the case of educational institutions is no different. Because of the presence of huge amount of structured data like the grades , enrollment data , progression rates as well as unstructured data like student opinions expressed through surveys , web blogs , twitter , Facebook etc., it becomes highly time and resource consuming to summarize the information manually to reach data led conclusions and decisions. Learning analytics (LA) is a field of data analytics that measures, analyses, reports and predicts data about learners for the purpose of optimizing teaching and learning. The data that gets normally analyzed are structured data including grades, attendance data, login frequency and site participation with respect to a Learning Management System (LMS) and so on. However, as pointed out by [3], [4] and [5] majority of the LA systems still do not capture unstructured data analysis leading to a user experience modeling to determine whether a learner is satisfied with the learning experience. On the other hand it is crucial to understand the patterns generated by the data like student feedback to effectively improve the performance of the institution and to create plans to enhance institutions' teaching and learning experience. From among the different mechanisms for collecting student feedback, surveys have an important role and most of the educational institutions undertake surveys in various forms. In Middle East College (MEC), to know the level of student satisfaction, in a year, three surveys are conducted viz. Student Satisfaction Survey, Module Evaluation Survey and Blitz Survey through which students give their opinion about various factors related to teaching and learning at the institution. The module evaluation survey and student satisfaction survey are conducted electronically whereas the Blitz survey is implemented manually. In all these three cases, the data analysis is done manually which causes substantial delays in taking appropriate decisions for improvement based on student concerns which results in less students satisfaction and less intake. To avoid this circumstance and to increase the revenue of the college, the proposed research is undertaken. The recent developments in data mining and analytics has brought together computational linguistics, natural language processing and machine learning together unleashing great prospects for organizations in automating analytics of unstructured text data. Opinion mining (OM) is one such advancement in the area of text mining which can be used primarily to determine the polarity of opinion from the vast set of dataset involving unstructured text data. This method has been used by the researchers in classifying opinions in various applications like e-commerce, movie reviews, product reviews etc. [6].

This paper focuses on using Opinion Mining technique for classifying the students' feedback obtained during module evaluation survey that is conducted every semester to know the feedback of students with respect to various features of teaching and learning such as module, teaching, assessments, etc. The extracted and preprocessed datasets were subjected to various supervised opinion mining algorithm such as Support Vector Machine (SVM), Naïve Bayes (NB), K Nearest Neighbor (KNN) and Neural Networks (NN) implemented using Rapid miner, the open source tool available for opinion

978-1-4673-9584-7/16/$31.00 ©2016 IEEE

classification. The comparative efficiency of the algorithms in the chosen application context is evaluated using precision, recall and accuracy measures.

Accuracy is defined as the ratio of total classifications that are precise to the total number of data set. Precision is the ratio of true positives to the total number of positives that are predicted whereas recall is the ratio of true positives with the total positives in the dataset as discussed in [20].

The rest of the paper is organized as follows. Section II reviews related work in the area of Learning analytics and Opinion Mining. Section III details the methodology of work undertaken for this paper. Section IV discusses the results and Section V provides conclusion and further research directions.

II. LITERATURE REVIEW

A. Opinion Mining

Opinion is the view of a person representing their sentiments, beliefs or judgments in regard to a matter of importance in a particular context and is normally considered to be subjective in nature. Studies show that more than facts, opinions of stakeholders greatly influence decision making of individuals as well as communities like governments and organizations.Opinion mining and sentiment analysis, the terms that are used interchangeably these days is a field of text data mining that involves extraction of opinions from evaluative texts and classification of the opinion's polarity as being positive or negative based on the orientation of the text results following the computational treatment of opinions expressed towards the key features[7].Since opinions are expressed in human language, Natural Language Processing(NLP) techniques are mostly employed in conjunction with KDD methods for various stages of opinion mining like opinioned statement detection, feature identification, opinion extraction , polarity determination and opinion summarization. From among the lexicon based approaches and machine learning approaches, supervised machine learning techniques based on algorithms like Support Vector Machine (SVM), Naïve Bayes (NB), K Nearest Neighbor (KNN) and Maximum Entropy etc. that uses large number of labeled training data are commonly employed to determine polarity for the purpose of classification [8].

The survey paper [7] aptly explains SVM, NB, NN and KNN classifiers as follows: SVM classifier works best for classifying sparse text data by defining rectilinear partitions in the data set and divides the set into different classes. The best partition plane is determined by the maximum normal distance between the data sets. The NB classifier is the most commonly used text mining classifier which uses Bayes theorem to calculate the possibility of the given label belong to a particular feature, p(l/f) using the below formula

$$P(l/f) = (P(l) * P(f/l))/P(f) \quad (1)$$

Where P(l) is the possibility of occurrence of a label in the dataset, P(f/l) is the possibility that a given feature belongs to a particular label. P(f) is the occurrence of a particular feature in the data set. If the features such as f1 ,f2, f3 …fn are not dependent on one another then equation (1) becomes

$$P(l/f)=(P(l) * P(f1/l)*P(f2/l)\ldots\ldots\ldots\ldots P(fn/l))/P(f)$$

Neural Network classifier employs multiple layer of neurons as a medium of classification where each neuron takes the word frequencies of the dataset as input. They are also associated with a weight for calculation of its input function. The output of each layer of neuron is back propagated to its other layers as training mechanism. The classifier predicts purely based on the input set, the weight and the trained neurons.

KNN classifier employs an indexing mechanism for the training data sets. To classify a document it calculates the similarity of the document with the training set index and uses the k-nearest by measuring the similarity by functions such as Euclidean distance.

Final results of OM heavily depend on the preprocessing or preparing the data before classification, representation of the text suitable for classification and the classifier used .Main tasks involved in data preprocessing are tokenizing – separating the sentences into words, removal of stop words-prepositions, and pro-nouns, etc. that does not give any additional meaning to the documents, stemming – converting various grammatical forms of words into root word, and generating n grams [9].References [10] and [11] point out that identification of the appropriate stop words to be removed has an impact on the quality of final classification.

Feature based opinion mining is yet another dimension that has been analyzed in the works of [12][13][14]

Opinion mining is being used now in retail industry for product reviews and recommender systems, service industry like education, health and tourism, governmental sectors – public opinion on policies, taxes, candidates and entertainment sector such as movie reviews etc. Opinion mining has been used to evaluate and classify student feedback from SMS as discussed in [15]. The author has developed three models, the base model with the necessary operators for classification, the second level model for data preprocessing and the last for performing sentiment analysis involving reading text resources, parsing SMS texts and categorizing text containing students' feedback.

The works that are mentioned in [16][17][18] also show how Sentiment Analysis have been performed on Student feedback and highlights the gaining popularity of this technique in educational sector. Reference [13] proposed a natural language processing technique using tool called GATE using common language engineering data structures and algorithms. The tool also includes set of components for information extraction and edition together with data visualization tools. The tool uses ANNIE technique which is used to create RDF or OWL (metadata) for unstructured content (semantic annotation) to polarize the parents and students feedback of any institution into positive, negative and neutral.

The research work detailed in [14] which combines data mining with natural language processing on a massive volume of data involving student feedback involves a comparative performance study of association rule mining and sequential pattern mining algorithms namely Apriori and Generalized

Sequential Pattern (GSP) mining in the context of extracting frequent features and opinion words using WEKA 3.7.10 a tool used for machine learning. The research also proved that GSP is more powerful in text mining than apriori.

III. METHODOLOGY

We have used Rapid Miner tool to mine the student feedback and classify it as positive and negative. The following methodology was adopted.

Step1. Conduct module evaluation and blitz survey

Step2. Preprocess the survey data

Step3. Extract entities and features to find their individual polarity

Step4. Train the model using SVM algorithm to classify into positive and negative classes

Step5. Test the model to determine accuracy.

Step6. Repeat steps 4 and 5 using K-NN, NB and NN algorithms

Step7. Compare the results of the performance for the four algorithms.

A. Module Evaluation Survey and Blitz Survey

The data set for the research was student feedback from six programs generated out of the Module Evaluation Survey (MES) that is conducted every semester in the college using the tool Survey Monkey and Blitz Survey that is conducted manually. Along with demographic questions and rating questions the surveys administer open ended questions that give the option to the respondents to provide detailed comments on different aspects of teaching and learning at the institution from which the data set for experimentation was primarily extracted. The Survey Monkey tool automatically generates the data in the form of MS Excel sheets and The Blitz survey results were manually tabulated into MS Excel files. Table 1 below presents the details of the full data set. The details of the dataset per program are presented in Table 2.

TABLE 1 SUMMARY OF DATASET

Total Number of responses overall	6433
Total Number of Comments:	12866
Average number of comments per response	2
Total Number of tokens	154335
Average number of tokens per response	24

TABLE 2 SUMMARY OF DATASET PER PROGRAM

Program Name	No: of Responses	No: of comments	No: of Tokens
CE	371	754	18064
Com	2810	5596	61799
ECE	1132	2263	26541
CFS	550	1134	17996
Math	808	1607	18648
MS	762	1512	11287

B. Data Preprocessing

The datasheet contained comments expressed in Arabic language and English Language. For this study, only the English responses were extracted and each response was stored as a separate text file. The positive and negative response files were grouped into separate folders which were later used as labels for training the supervised learning algorithms experimented. The response statements were taken through preprocessing steps of tokenization, stop words removal and stemming using appropriate operators in Rapid Miner. Word vector representations were also generated using the TFIDF method. All the aforementioned processes of data preprocessing and word vector generation were achieved using the *ProcessDocumentsFromFiles* operator in Rapid Miner which can contain all the other required operators. Fig.1. depicts the main process that contains all the operators including *ProcessDocumentsFromFiles* which takes as input positive and negative folders that contain text files.

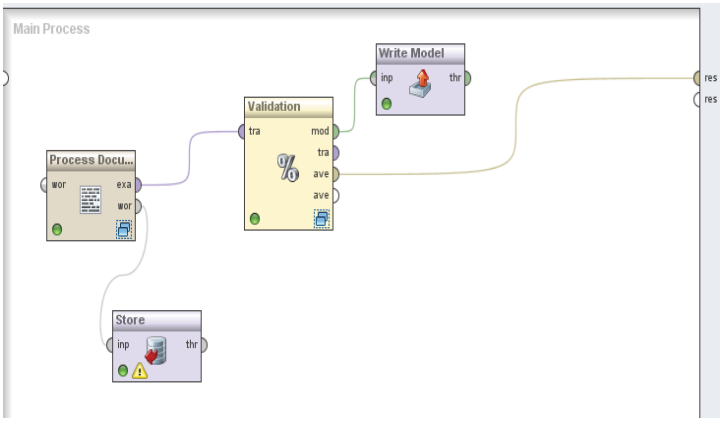

Fig.1. Main Process.

C. Entity and Feature Extraction

In order to understand the polarity of opinion with respect to various features, the terms Module, Teacher, Exam, Resources were chosen. These words and their synonyms were applied to the operators *SelectAttributes* and *FilterRows* in Rapid Miner to understand the polarity feature wise.

D. Training and Validating the models

The supervised learning algorithms that were used are SVM, NB , K-NN and NN .The Validation operator from Rapid Miner which allows to simultaneously train and test the classifiers was employed .In particular, the facility of cross validation was exploited for the input dataset by setting the value to 10. This meant that the data set was divided into 10 groups of which first 9 groups were used as training and one group was used for testing. In the second run a different combination of 9 sets became the training data and a new set became the testing data. The process was continued until all the permutations will be finished.

E. Comparison of results

To compare the performance of the four algorithms employed, Accuracy, Precision and Recall values were calculated for

each of the classifier algorithm by using the Performance Operator of Rapid Miner. Training, Testing and Performance evaluation process as undertaken is portrayed in Fig. 2.

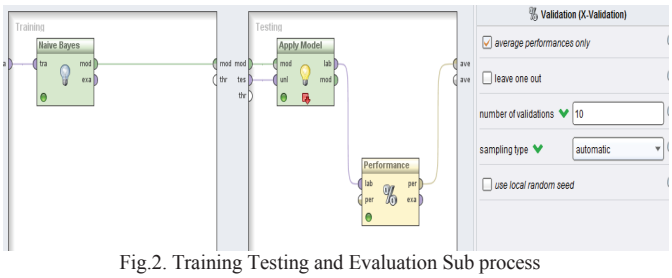

Fig.2. Training Testing and Evaluation Sub process

IV. FINDINGS

TABLE 3 PERFORMANCE COMPARISONS OF SVM and K-NN

Program	SVM			K-NN		
	P	R	A	P	R	A
CE	100	30	85.97	100	85	97.78
Comp	100	93.65	95.84	100	97.69	98.45
ECE	100	81.67	91.54	100	95	97.69
CFS	100	70	92.56	100	100	100
Math	100	63.3	90.4	100	93.33	97.89
MS	75	100	92.18	100	90.83	97
Avg	95.83	73.10	91.41	100	93.64	98.13

TABLE 4 PERFORMANCE COMPARISONS OF NB and NN

Program	NB			NN		
	P	R	A	P	R	A
CE	100	100	100	100	95	98.75
Comp	98.5	98.4	98.8	100	99.21	99.49
ECE	100	100	100	100	100	100
CFS	100	90	97.89	100	91.67	97.81
Math	100	100	100	100	100	100
MS	100	94	98	92.31	93.94	95.10
Avg	99.75	97.07	99.11	98.71	96.67	98.52

Based on the values of table 3 and 4, the graph in fig 3 is plotted which shows the performance of SVM, KNN, NB and NN for all the 6 programs. The Result shows that KNN shows the best precision result of 100%, NB gives the best recall and accuracy result of 97.07% and 99.11% respectively.

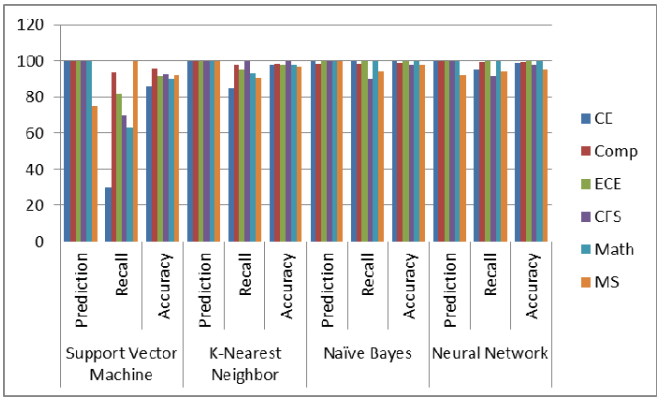

Fig.3. Performance comparison results of SVM, KNN, NB and NN

Table 5 shows the results of Opinion classification based on extracted features such as teacher, exam, module content and Resources/lab resources. The result shows less positive results as students are encouraged to give negative feedback in module evaluation survey whereas in blitz survey conducted in the beginning of every semester, students are encouraged to give both positive to negative feedback. The result shown in fig.4 infers that the feature exam needs more improvement as it fetched the most negative comments.

TABLE 5 FEATURE BASED OPINION EXTRACTION AND CLASSIFICATION

Feature	True Positive	True Negative
Teacher	87	128
Exam	69	267
Module Content	15	54
Resources/Lab Resources	66	263
Average	59.25	178

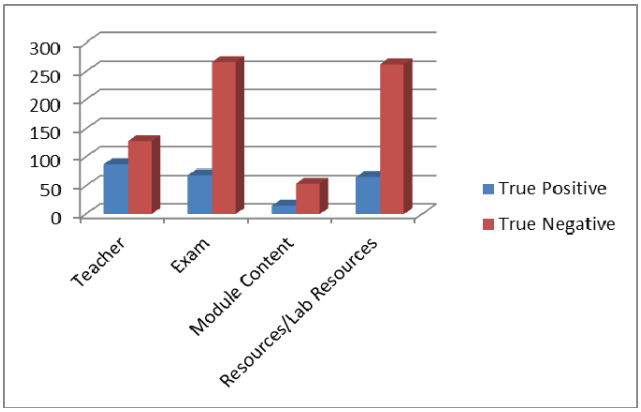

Fig.4. Feature Based Opinion Classification

V. CONCLUSION

In this research, we performed opinion mining on the student feedback generated through surveys using supervised machine learning algorithms implemented through Rapid Miner. Our results indicated that Naïve Bayes algorithm outperformed others in terms of accuracy and recall and K-Nearest Neighbor algorithm performed the best in terms of precision. Similarly key features of teaching and learning at Middle East College is extracted from the students' feedback and classified to be positive and negative based on their polarity to analyze the feature which needs improvement. In future research, we intend to perform opinion mining of student feedback gathered using social media and also to comparatively analyze how the student opinion varies using various demographics like age gender and so on. We are also planning to improve the performance of opinion mining process by using map reduce framework by which program wise , year wise analysis can be run parallel.

ACKNOWLEDGMENT

The authors would like to acknowledge the support and guidance provided by senior management of Middle East College for providing necessary infrastructure and training in carrying out this work.

REFERENCES

1. Khan, Khairullah, "Mining opinion componentsfrom unstructured reviews: A review," Journal of King Saud University – Computer and Information Sciences , Vol. 26, 2014.

2. Jyotsna Talreja Wassan, "Discovering Big Data Modelling for Educational World", IETC Procedia - Social and Behavioral Sciences, pp:642 – 649, 2015

3. Dirk T. Tempelaar, "In search for the most informative data for feedback generation: Learning analytics in a data-rich context", :Journal of Computers in Human Behavior, Vol 47, pp: 157–167, 2015.

4. Beth Dietz-Uhler and Janet E. Hurn, "Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective", Journal of Interactive Online Learning, Volume 12, Number 1, 2013.

5. Marie Bienkowski, Mingyu Feng, "Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics", Department of Education, Office of Educational Technology: October 2012.

6. Pang, B. and Lillian L. s.l, "Opinion mining and sentiment analysis : Foundations and trends in information retrieval", Vol. 2, 2008.

7. Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, Vol. 5, 1093–1113, 2014.

8. Kumar Ravi, Vadlamani Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", Published in Knowledge-Based Systems Vol. 89 , 14–46, 2015.

9. S.Vijayarani, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, Vol 5(1),7-16.

10. Munkova, Dasa, Munk, Michal and Mozar, Martin, "Data Pre-Processing Evaluation for Text Mining:Transaction/Sequence Model", Procedia Computer Science, Vol. 18, 2013.

11. Ramasubramanian and Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.

12. Gautami Tripathi and Naganna, "Feature selection and classification approach for sentiment analysis", Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015.

13. Trisha Patel, "Sentiment Analysis of Parents Feedback for Educational Institutes", International Journal of Innovative and Emerging Research in Engineering, Volume 2, Issue 3, 2015.

14. Padmapani and Tribhuvan, "A Peer Review of Feature Based Opinion Mining and Summarization", International Journal of Computer Science and Information Technologies, Vol. 5 (1), 247-250, 2014.

15. Chee Kian Leong, "Mining sentiments in SMS texts for teaching evaluation", Journal of Expert Systems with Applications, Vol. 39, 2584–2589, 2012.

16. Chien-wen Shen, "Learning in massive open online courses: Evidence from social media mining", Journal of Computers in Human Behavior, vol. 51 568–577, 2015.

17. Sunghwan Mac Kim and Rafael, "Sentiment Analysis in Student Experiences of Learning", Available at ResearchGate.com

18. Alaa El-Halees, "Mining Opinions in User-Generated Contents to Improve Course Evaluation", Communications in Computer and Information science, January 2011.

19. Dietz-Uhler, Beth and Hurn, "Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective", Journal of Interactive Online Learning, Vol. 12, 2013.

20. Rehab M. Duwairi, "Arabic Sentiment Analysis using Supervised Classification", "The 1st International Workshop on Social Networks Analysis, Management and Security (SNAMS - 2014), August 2014.

21. Tempelaar, Dirk, Rienties, Bart and Giesbers, Bas, "In search for the most informative data for feedback generation:Learning analytics in a data-rich context", Computers in Human Behavior, Vol. 47, 2015.

22. Vijayarani, Ilamathi, and Nithya "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, Vol. 5.

23. Tan, Ah-Hwee. "Text mining: The state of the art and the challenges" Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, Vol. 8, 1999