Mining Unit Feedback to Explore Students' Learning Experiences

Zainab.Mutlaq Ibrahim, Mohamed Bader-El-Den, Mihaela Cocea
(zainab.mutlaq-ibrahim, mohamed.bader, mihaela.cocea)@port.ac.uk

University of Portsmouth,Lion Terrace PO1 3HE,UK

Abstract: Students' textual feedback holds useful information about their learning experience, it can include information about teaching methods, assessment design, facilities, and other aspects of teaching. Analyzing such feedback can form a key point for educators and decision makers to help them in advancing their systems. In this paper, we proposed a data mining framework for analyzing end of unit general textual feedback using four machine learning algorithms, support vector machines, decision tree, random forest, and naive bays. We filtered the whole data set into two subsets, one subset is tailored to assessment practices(assessment related), and the other one is the non-assessment related data subset, We ran the above algorithms on the whole data set, and the new data subsets. we also, adopted a semi automatic approach to check the classification accuracy of assessment related instances under the whole data set model. We found that the accuracy of general feedback data set models were higher than the accuracy of the assessment related models and nearly the same value of the non- assessment related modeles. The accuracy of assessment related models were approximated to the accuracy of the assessment related instances under the full data set models.
Keywords: Sentiment Analysis,Educational Data Mining, Assessment, Student Feedback.

1 Introduction

Unit feedback is a fundamental part of the learning process for institutions. It can provide data about units in which may hold implicit useful knowledge for researchers and practitioners to understand student learning experiences. This can form a start point for educators to modify or develop units accordingly.
Normally, feedback consist of a simple survey form, most often a combination of Likert scale responses to questions or statements(Responses can be strongly agree, agree,neutral, disagree or strongly disagree), in addition to one or more of open-ended question(s) where students need to write few short sentences.
Quantitative data can be taken from likert scale responses. In fact, these responses have been used for long time to assess the teaching effectiveness[1], due to the fact that they are numerical data and easy to analyse, on the contrary of the Open-ended responses which are not easy to analyse as likert responses. Open ended responses are unstructured data, also they may have keywords that are not included in Likert questions' words or even in their own words in which

make them (the open-ended responses) not bias to the trend of a survey.They can be a very important source of any faculty analytic processes to reveal the hidden information in these texts.And they can be a big chance for students to be a real participant of the ongoing studies to advance education.
students' feedback have been to some extent accepted by researchers and practitioners due to the fact that student ratings and feedback are the most valid source for evaluating teaching effectiveness [2], students are the people who receive the teaching procedures, so their feedback is crucial.
Educational data mining (EDM) is the discipline that focuses on developing methods and algorithms to explore big data that comes from educational databases and sources to better understanding of students and the setting they learn in [3].In particular, sentiment analysis is the process of analyzing statements and obtaining subjective information from them.
Massive online open courses(MOOC) have been accepted by some institutions since 2012, with such a huge class size, massive blog feedback is being generated which form a big challenge for EDM community to innovate and advance frameworks and techniques to analyze such feedback, hence, this framework can be used to analyse such data.
The rest of the paper is organized as follows: Related work is presented in section 2. The framework and work flow is presented in section 3. The used data set of this study is presented in section 4. Experiments and results are presented in section 5, and finally conclusion and future work are presented in section 6.

2 Related Work

Researchers have been encouraged and motivated by EDM community to innovate new frameworks to analyze different educational topics such as assessment,students' emotion,browsing or interaction data, the results of educational research, and many more [4]. EDM community also has urged researchers to apply a previously used frameworks to a new domain or reanalyze an existing data set with a new technique [4].
Data mmining algorithims have been applied, to classify students according to their Moodle usage and the final obtained marks[5], to predict student retention [6], to reduce dropout rates[7], to analyze students' programming assignment [8], and many more.
All of the above studies [5,6,7]used a structured data sets which were taken from Moodles and databases, structured data sets is easy to search by simple algorithm , example of this is spread sheets, while unstructured data such as tweet is more like human language and searching it is very difficult and needs an advanced and special algorithms .
Analyzing unstructured data(text mining) needs a pre-processing stage to structure it so it becomes easily search-able and manageable. Recently,there is an increasing number of research in utilizing text mining techniques for different educational purposes and applications due to the need of advancing and developing learning process. Abd-Elrahman et al [9]used WordStat tool to determine

the number of positive and negative entries with different teaching aspect categories, they mentioned only spelling errors and nothing about the dimensions of the data, they depends on a algorithm that count the number of occurrence of a specific words. Sliusarenko [10] transformed the qualitative feedback into quantitative by extracting the key-terms,then applied factor analysis to find the most important factors in the feedback, and finally applied regression technique to see which factors have the most impact on student ratings. Pan et al. [11] SPSS text analysis in a try to strengthen quantitative data with systematic and meaningful qualitative interpretation. Jordan [12] utilized StatSoft Statistica and SPSS, he built a correlation model using text mining methods,he found a weak correlation between the Likert responses and the open-ended written responses. This means there are significant words and patterns within the open-ended responses that can provide additional information to the decision makers. Finally, Pagare,Chen et al. [13,14] analyzed twitter data to understand students' learning experiences.They[14] innovated a new work flow which was a mixed of human efforts and machine learning. Both [14,13] applied Naive Bays(NB). Chen also applied Support Vector Machine(SVM)and Max Margin Multi-label (M3L) classifiers [14].They [14,13] concluded that NB is very good classifier to use on text data.

3 Framework

In this section, we present a general framework for analyzing end of unit students' feedback which was given in text format as shown in Figure.1.

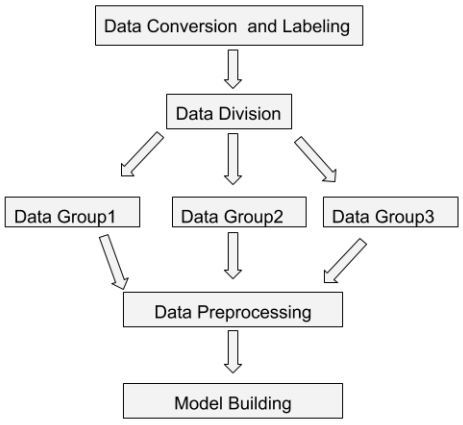

Fig. 1. Proposed general framework of students' feedback model

The main aim of this study is to develop a data mining framework to capture students' concern of assessment from general feedback and to investigate the performance of data mining models when they are applied to general feedback verses topic specific. To achieve the above aim, the following objectives are defined:

- Instances classification:Identify the best classification model to automatically detect the class of each instance, to filter and divide the full data set into assessment and non-assessment related sub sets.
- Assessment related instances' sentiment: Identify the best sentiment analysis model that automatically detect the polarity of the assessment related instances, so we can identify issues from negative instances.
- Full data set instances' sentiment: Identify the best sentiment analysis model that automatically detect the polarity of its instances.
- Assessment related instances' sentiment under the full data set model: Detect the polarity of the assessment related instances under the sentiment of whole data set model.
- Non-Assessment related instances' sentiment: Identify the best sentiment analysis model that automatically detect the polarity of the non-assessment related instances.

4 Data

The used dataset in this study is a hand-written text, it was collected from students as end of unit(INDADD) feedback for the years 2012-2016, it consists of 979 instances, it includes responses to the following two Statements:
Statement1: The best part of the unit is: , Statement2:The area of this unit that needs improving is:, Statement1 considered as positive and statement2 as negative feedback.

4.1 Labeling

The data has been labeled using three labels: Assessment related label, where some keywords and their derivatives or synonymous are presented,such as "coursework", "exam", and "quiz", "assessment", "marking", "grading","test", "feedback" and "evaluation". Assessment not related label, where there is a meaningful text but the assessment related keywords mentioned above are not presented. And irrelevant label to cover the empty instances, misspellings, jokes, or irrelevant statements which we have none of them in our data set.
Three native English speakers reviewed the feedback data and label it according to the suggested labels. Each entry of the feedback has three rating from the three viewers. Although, rules of labeling are very clear, easy to follow, and far to mislead, the raters have a chance to label the entry as irrelevant.
In content analysis literature, statistical measures such as scott's Pi, Fleiss Kappa and Krippendorf's Alpha are used to decide agreement among raters on

topics[15,16].However, Chen and his colleagues [1] used the harmonic mean(F1 measure) to measure how close two label sets are assigned to one entry by two raters as their study were dealing with multi-label classification problem. In our study, the defined labels are mutually exclusive which means that each entry can fall under only one label , so any one of the above measures is applicable to our data set.
In this study Krippendorf's Alpha measure is adopted as a reliability coefficient, it was developed to measure the agreement among raters on a specific topic , it applies to all sort of metrics,any number of raters, any number of categories,incomplete data,and big or small data samples[17].
As this study utilizes Krippendorf's Alpha measure which effectively can deal with the missing data[17], the raters above can have extra chance to skip labeling in tricky text. The Krippendorff's score is 0.9841 which considered to be an optimist value.

4.2 Data Distribution

The used dataset in this study is anonymous data, which means that students can not be identified but sometimes the contents of the data identified few lecturers, however to make the contents anonymous for the lecturers too, we refer to the persons by index such as L1(for the first mentioned lecturer),L2, ... and so on. see Figure. 2.

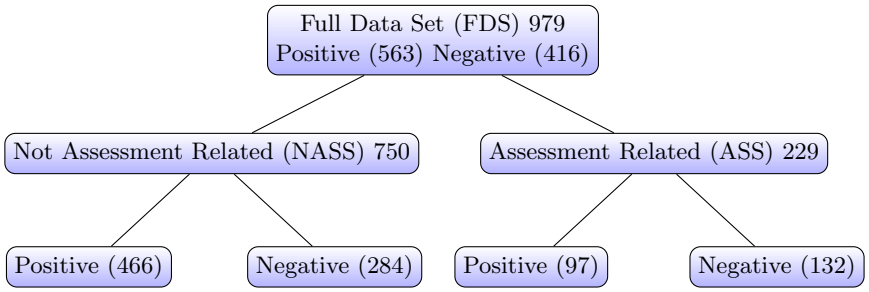

Fig. 2. Data Set Distribution

5 Experiments and Results

In this section, we executed five experiments to fulfill the objectives that which are mentioned in the framework section. For each experiment, we used ten fold cross validation method as this method is robust against potential bias to the training data set, in this method we used nine fold for training and one fold for testing, we repeated the test ten times and calculated the average performance for all attempts. We used a PC desktop with with quad 2.33 GHZ CPU, 4GB

RAM, and windows 10 operating system.The following section includes the pre-processing components that were executed to all experiments. Followed by a brief description of classifiers that we are going to use and four popular evaluation measures.

5.1 Text Pre-Processing

Hand-written feedback can include miss-spelling errors, jokes, irrelevant statements, some special characters such as "@", Punctuation marks, etc. However, it is the open-ended nature of questions that allows students to express what is in their minds and what they feel without the constraint of the carefully worded numerical rating questions.
Pre-processing stage aims at cleaning the data and reducing its dimension, this can contribute positively to more accurate results. In this section the following operations are executed using KNIME analytic platform :

- Punctuation Erasure: Removes all punctuation characters of terms contained in the text, such as exclamation, question marks.
- N Chars Filter: Filters all terms contained in the text data with less than the three characters such as"@","in","all", "the", and many more.
- Number Filter: Filters all terms contained in the text data that consist of digits, including decimal separators "," or "." and possible leading "+" or "-".
- Case converter: Converts all terms contained in the text data to lower.
- Stop word Filter: Filters all terms of the the text, which are contained stop word such as "because", "again", and "the".
- Snowball Stemmer: Stems terms contained in the text data with the Snowball stemming library to guarantee that each term represent once and only once in the created bag of words (BoW).
- Feature creation: In text mining, text's features are the characters or the words of that text. Feature selection is the process of eliminating the irrelevant and trivial features and keeping the significant features which are genuinely affecting the performance of the constructed model. To explain that more let us have this example from our data set, "That it was 100 percent coursework based on real life seminars which we could easily relate to" , in our study which is about assessment ,the most significant feature in the above example is the "coursework" term , however, in machine learning we can not eliminate all features and keep the most significant one.
  Some of the text pre-processing operations contribute to feature selection process, such as removing stop words, number filter, snowball stemmer, and n-character filter.
  One of the most popular feature creation technique is n-grams[18,19]. An n-gram is a sere of n items from a text, it can be letters or words,uni gram,is very popular technique which is selecting single words, bigram is selecting two words at a time for example "the coursework is not clear", "the

coursework", "coursework is", "is not", "not clear" . In sentiment analysis section(see table 2), we are going to use uni gram(UNIGRAM) and bi gram(BIGRAM) .

5.2 Classifiers

The most popular classifiers for text mining are: Support Vector Machines(SVM) [20,21,22], it is a powerful tool for solving data mining problems such as classification, regression, and feature selection, it has the power to determine an optimal separating point that labels records into different categories [23]; Naive Bayes [1,13,20,24,22], is a probabilistic classifier, it is robust to noisy data and irrelevant attributes, and can cope very well with null values[25]; Decision Tree[24,22], it uses training examples of data to construct the tree, at classification time the tree executed from root to leaf, so the leaf node decides the class of the record [25]; and Random Forest[24,22], it selects attributes randomly and utilizes the decision tree as the base model [25].

5.3 Assessment Classification

Experiment1:
The first experiment aims to build a model that automatically filter the assessment related instances from the whole data set,we labeled the data as mentioned in labeling section and ran the text pre-processing component to build the final model.
Table 1 illustrates the evaluation results of experiment 1 performance.

Table 1. Experiment 1

	NB	SVM	DT	RF
Accuracy	0.9640	0.9930	1	0.7640
Precision	0.9770	0.9855	1	0.6725
Recall	0.9255	0.9950	1	0.5450
F-score	0.9480	0.9950	1	0.5295

From the result, we observed that all classifiers' performance in terms of accuracy, precision,recall, and f-score was significant except of the RF classifier which performed poorly.The optimal classifier was DT as it was error free.

5.4 Sentiment Analysis

In this section we executed four sentiment analysis experiments(2, 3, 4, 5), Figure 3 shows their proposed framework. Table 2 illustrates the result of Experiments(2,3,4,5) using uni-gram(UNIGRAM) and bi-gram(BIGRAM) features,only experiment 5 used the results of experiment 3 using uni-gram.

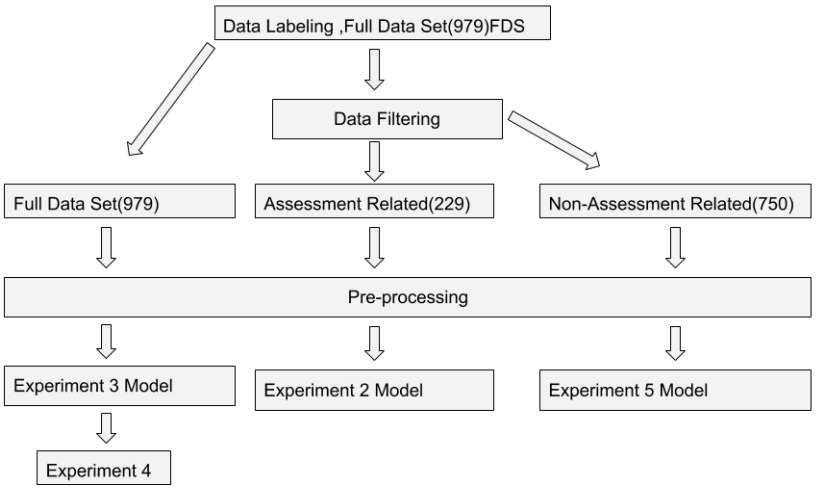

Fig. 3. Proposed framework for experiments: 2,3,4,5

Sentiment Analysis of Assessment Related Instances
Experiment2:
In this experiment we used 229 of assessment related instances(ASS) to build a model that automatically detect their sentiment. SVM classifier was the best performer in terms of accuracy , it was 72 % percent accurate.The poorest performer was the RF classifier, as its error rate was 42% . The recall values show that SVM is the most sensitive of the four classifiers, i.e, it correctly identifies instances of both classes,while RF is the least sensitive.Precision is the highest for SVM and lowest for RF. The best balance between precision and recall is attained by RF which is only 1% higher than the value that attained by SVM.

Sentiment Analysis of Full Data Set Instances
Experiment3:
We used 979 instances to build a model that automatically detect the sentiment of the whole data set.Also, SVM classifier recorded the higher accuracy of 76% followed by DT,then RF. The poorest performer classifier is NB with error rate of about 40 %,this opposed to the finding of Chen and Pagare[1,13] that NB is a very good for text classification. The recall values show that SVM is the most sensitive, it has the highest value of correctly identified instances of both classes,while NB is the least sensitive.The highest precision is achieved by SNM, while the lowest is achieved by NB.The best balance between precision and recall is achieved by SVM,while the lowest balance is achieved by NB.

Sentiment Analysis of Assessment Related Instances under the Full Data Set Model

Experiment4:
This experiment was built on the results of experiment 3 which includes a total of 979 instances. The aim was to view how accurate the general feedback(FDS) model in classifying the assessment related instances(ASS). The highest accuracy was scored by RF model, while the least value was scored by NB.

Sentiment Analysis of Non- Assessment Related Instances
Experiment5:
We used the 750 of non-assessment related(NASS) instances to build a model that automatically detect their sentiment. SVM outperforms all other classifiers in terms of accuracy, precision, recall, and the F-score.

Table 2. Results of experiment:2,3,4,5

			Accuracy	Prcision	Recall	F-score
DT	Exp2	ASS-UNIGRAM	0.67	0.67	0.67	0.67
		ASS-BIGRAM	0.65	0.64	0.64	0.64
	Exp3	FDS-UNIGRAM	0.71	0.70	0.69	0.70
		FDS-BIGRAM	0.71	0.70	0.69	0.70
	Exp4	ASS-IN-FDS	0.66	–	–	–
	Exp5	NASS-UNIGRAM	0.72	0.70	0.69	0.69
		NASS-BIGRAM	0.71	0.69	0.69	0.69
SVM	Exp2	ASS-UNIGRAM	0.69	0.70	0.70	0.69
		ASS-BIGRAM	0.72	0.72	0.73	0.72
	Exp3	FDS-UNIGRAM	0.76	0.76	0.74	0.74
		FDS-BIGRAM	0.76	0.76	0.74	0.76
	Exp4	ASS-IN	0.69	–	–	–
	Exp4	NASS-UNIGRAM	0.76	0.75	0.73	0.74
		NASS-BIGRAM	0.75	0.74	0.72	0.73
RF	Exp2	ASS-UNIGRAM	0.58	0.58	0.50	0.73
		ASS-BIGRAM	.058	0.59	0.51	0.38
	Exp3	FDS-UNIGRAM	0.73	0.75	0.71	0.71
		FDS-BIGRAM	0.74	0.77	0.61	0.67
	Exp4	ASS-IN-FDS	0.70	–	–	–
	Exp5	NASS-UNIGRAM	0.62	0.31	0.50	0.38
		NASS-BIGRAM	0.63	0.31	0.50	0.38
NB	Exp2	ASS-UNIGRAM	0.66	0.68	0.62	0.60
		ASS-BIGRAM	0.67	0.73	0.63	0.60
	Exp3	FDS-UNIGRAM	0.60	0.59	0.56	0.55
		FDS-BIGRAM	0.61	0.60	0.57	0.55
	Exp4	ASS-IN-FDS	0.44	–	–	–
	Exp5	NASS-UNIGRAM	0.64	0.61	0.58	0.57
		NASS-BIGRAM	0.65	0.62	0.57	0.57

5.5 Comparison of Experiment 2,3,4,5 Results

We observed that in all experiments there was no significant difference between using a uni-gram and bi-gram features.

Although ASS and NASS are subsets of FDS, there is a notable margin between the accuracy values of FDS models and ASS models, but this did not apply to the NASS models as its accuracy value is approximated to the accuracy value of FDS models. For example, the accuracy of SVM for FDS models and NASS models are nearly the same, while it is different from ASS models by 7%. This applies to DT models, but not to RF and NB models.
The accuracy values of all ASS models are very close to the accuracy values of assessment related instances under FDS models which means it is better to apply classifiers directly on topic domain data set.

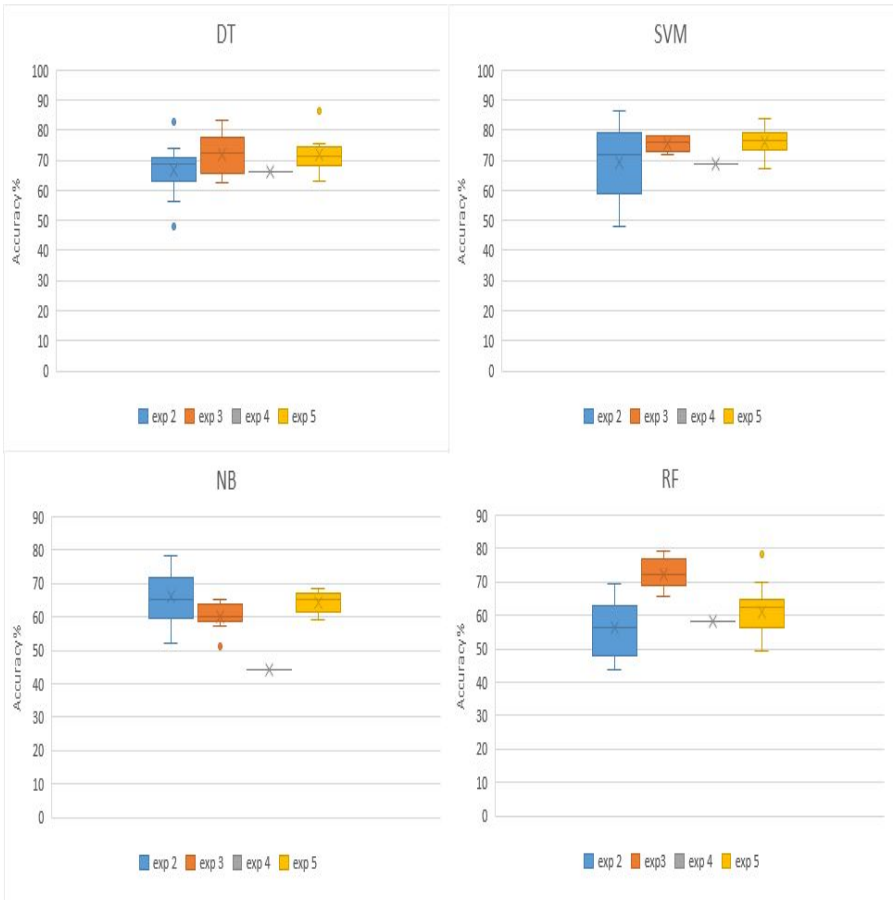

Fig. 4. Accuracy range for experiments: 2,3,4,5

6 Conclusion and Future Work

In this paper we examined the learning potential of four machine learning classifiers for learning classification and sentiment from students' textual feedback. We ran SVM, DT, NB, and RF classifiers on our data set and its subsets, the FDS was 979 instances, ASS subset was 229,and NASS subset was 750.
Our experiments showed no significant difference between using unigram and bigram features in building our models.
We evaluated each classifier performance. DT classifier was error free for Experiment 1,SVM outperformed DT, NB, and RF in all sentiment analysis experiments 2, 3 , 4, 5. Figure 4 illustrates the accuracy range and average of all models ,we found that the accuracy of general feedback data set machine learning models were higher than the accuracy of the assessment related machine learning models and nearly the same value of the non- assessment related machine learning models. The accuracy of assessment related machine learning models were approximated to the accuracy of the assessment related instances under the full data set machine learning models.
We used different parameters in each model, bias and gamma in SVM models, default probability in NB models, a static random seed and Gini index in RF models, and attribute selection and no pruning in DT models.
Future work includes more, analyzing of students' sentiment of assessment,recognizing the assessment issues, and using part of speech feature(POS) in building our models.

References

1. Howsher L Chen Y. Student evaluation of teaching effectivness:An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education*, 2003.
2. Schmelkin L. P. Spencer K.J. Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education*, 27(5):397–409, 2002.
3. Educational Data Mining deffenetion. Accessed: 24-10-2017.
4. Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.
5. Ventura Sebastian Espejo Pedro G Hervas Cesar Romero, Cristobal. Data mining algorithms to classify students. *Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings.*, 2008.
6. Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal. Mining education data to predict student's retention: A comparative study. *arXiv preprint arXiv:1203.2987*, 2012.
7. Saurabh Pal. Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business*, 4(2):1, 2012.
8. E. Albrecht and J. Grabowski. Towards a framework for mining students' programming assignments. pages 1096–1100, April 2016.
9. Abbott T Abd-Elrahman A, Andreu M. Using text data mining techniques for understanding free-style question answers in course evaluation forms. *Research in higher Education Journal*, 9:12–23, 2010.

10. Ersboll B K Sliusarenko T, Clemmensen L K H. Text mining in students' course evaluations relationships between open-ended comments and quantitative scores. *Proceedings of the 5th International Conference on Computer Supported Education*, pages 564–573, 2013.
11. Daphne Pan, Gary SH Tan, Kiruthika Ragupathi, Krishna Booluck, Rita Roop, and Yuen K Ip. Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. *Research in Higher Education*, 50(1):73–100, 2009.
12. D.W Jordan. *Re-thinking Student Written Comments in Course Evaluation: Text Mining Unstructured Data for Program and Institutional Assessment* . PhD thesis, California State University, 2011.
13. Pagare Pallavi. Recognizing student's problem using social media data. *International Journal of Computer Science and Mobile Computing*, 4(6):440–446, 2015.
14. Madhavan K Chen X, Vornoreanu M. Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, 7(3):246–259, 2014.
15. Krippendorff K. Reliability in content analysis. *Human Comm. Research*, 30(3):411–433, 2004.
16. Bracken c.c Lombard m, Snyder-Duch J. Content analysis in mass communication:assessment and reporting of intercoder reliability. *Human Comm. Research*, 28(4):587–604, 2006.
17. Krippendorff K. Computing Krippendorff's Alpha-Reliability. 2011.
18. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
19. Bhayani R. Go A., Huang L. Twitter Sentiment Classification using Distant Supervision. 2017.
20. Yelena Mejova. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*, 2009.
21. R de Groot. Data mining for tweet sentiment classification. Master's thesis, 2012.
22. Feng Tian, Pengda Gao, Longzhuang Li, Weizhan Zhang, Huijun Liang, Yanan Qian, and Ruomeng Zhao. Recognizing and regulating e-learners' emotions based on interactive chinese texts in e-learning systems. *Knowledge-Based Systems*, 55:148–164, 2014.
23. Pao HK Lee YJ., Yeh YR. Introduction to support vector machines and their applications in bankruptcy prognosis. 2012.
24. Filipe R Lucini, Flavio S Fogliatto, Giovani JC da Silveira, Jeruza L Neyeloff, Michel J Anzanello, Ricardo de S Kuchenbecker, and Beatriz D Schaan. Text mining approach to predict hospital admissions using early medical records from the emergency department. *International journal of medical informatics*, 100:1–8, 2017.
25. Du H. *Data Mining Techniques and Applications.* International series of monographs on physics. 2010.