Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 178 (2020) 194–204

Procedia
Computer Science

www.elsevier.com/locate/procedia

9th International Young Scientist Conference on Computational Science (YSC 2020)

A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments

Ochilbek Rakhmanov[a]*

[a]*Nile University of Nigeria, Abuja, Nigeria*

Abstract

Sentiment analysis is one of the important fields in educational data mining. In this paper, a large dataset, more than 52 000 comments, was used during experiment to develop a state-of-art classification model. The correlation test was conducted on sentiment analysis results and scale-rated survey results, and the result (r(203)=.79, p<.001) shows that sentiment analysis can be accepted as reasonable method for course and lecturer evaluation. A comparative analysis was done between different vectorization and classification techniques. The results of the experiment show that classifier built using Random Forest was most optimal and efficient classification model with state-of-art prediction accuracy of 97% for 3-class classification. Moreover, to improve the diversity of the comments, a 5-class dataset was formed and experiment resulted with an efficient classification model with accuracy of 92%. The Tf-Idf vectorization technique performed better than Count (Binary) vectorization.

Keywords: Sentiment analysis; vectorization; comment classification.

1. Introduction

Educational Data Mining is the research area developed to address issues in education using various tools of data mining techniques. Looking into patterns of data produced by students, lecturers or management can provided meaningful insights to improve the quality of education. Feedbacks from students to lecturers or from lecturers to students is vital part of this evaluation in education. The common methods of collecting feedbacks are filling scale-rated surveys and provide an open-end text comment. These feedbacks enable head of units and deans to make

* Corresponding author. Tel.: +2349090038311.
E-mail address: ochilbek.rakhmanov@nileuniversity.edu.ng

10.1016/j.procs.2020.11.021

objective evaluation of the course taught, hence, the outcomes, after evaluation, can serve as a good tool to improve the education quality [1]. While analyzing scale-rated surveys are relatively easy task, classification and finding patterns in text comments was always a difficult task. Thus, several methods of sentiment analysis were employed in many recent researches to address this issue [1–4].

Sentiment Analysis (SA) is a field that works on making sense out of textual feedback and opinions [2]. Opinions can vary in many different aspects, they can be negative or positive, and different emotions can be expressed like love, support or disappointment. Apart from it, one can provide a neutral opinion as well, not positive nor negative. Thus, extracting main features from text structures and classifying them accordingly is main job in SA task. SA was successfully employed before in many fields, like movie rating, service rating, product rating etc. [5] and recently gained popularity in the education field as well.

Many institutions used to evaluate student's satisfaction about the lecturer and the course through scale-rated surveys. One shortcoming of this system is that it is restricted, student cannot express extra feeling about lecturer or course. Thus, students were provided an opportunity to write a comment about their studies. Numerical scores (from scale-rated surveys) are taken as main indicator of student satisfaction, although much insight can be gained from analyzing the free text comments [6]. To overcome this issue, SA was employed successfully to extract meaningful and structured information from students comments in many researches [1–4].

There are currently two main approaches in SA, lexicon-based classification and machine learning based classification. Lexicon-based approach uses dictionaries to determine polarity of the text to classify them accordingly [5, 7]. Machine learning approaches differ from lexicon-based approaches, as they convert the textual data into numeric data, after several pre-processing operations, and uses classification algorithms to classify the data accordingly [1–3]. This paper concentrated on machine learning approach.

Researchers used various mediums to collect the students' comments. While previous studies suggested to collect feedback through Short Message System(SMS), later, most of the researchers concentrated on using of Student Response System (SRS) and social media [2]. While using social media is cheaper and easy method [2, 8], the authenticity of the comments is questionable, as everyone can comment, even non-participants, and same student can comment many times using different profiles. Thus, most convenient way is to use SRS provided by the university. At the end of each semester students are requested to fill a survey about each course and lecturer. This survey can also contain a section where a student can write open end comments. Later, these comments can be extracted from database. This approach was adopted by some of the recent researches and was also used during this research [1, 4, 6].

1.1. Challenges in educational sentiment analysis

There are several challenges need to be addressed in educational sentiment analysis. The very first concern is that none of previous studies addressed the issue of validation of sentiment analysis results. The used methods were just straight forward; to develop the classification model and check the prediction accuracy. As it was stated before, mostly the results of scale-rated surveys are being used to evaluate the course and lecturer's performance. In a same manner, sentiment analysis also contributes to the evaluation of the performance, but the question is this; how reliable are the scores of the comments which are given during sentiment analysis? Thus, a comparative study on sentiment analysis scores and scale-rated survey scores need to be addressed. Conducting such study would boost the trust to educational sentiment analysis methods.

Secondly, the dataset size was relatively small in many researches, thus, putting a question mark on developed models. Following are number of inputs in several recent researches: 101 comments [9], 200 comments [8], 1822 comments [1], 2254 comments [6], 3000 comments [10], 3800 comments [3] and 6000+ comments [4]. Due to, there is a need to conduct and experiment using machine learning techniques with big enough dataset, which was done in this research.

Almost all of literatures mentioned till now used simple count (binary) vectorization method during text-to-vector conversion. In brief, count vectorization assigns value "1" if the particular word presents in the sentence and "0" otherwise. But this bringing some shortcoming with itself, for instance, a word "chalk" has same weight with word "boring" while both of them carry completely different meanings which should affect the classification. Thus, employment of Tf-Idf weight assignment approach should be tested and compared to the performance of count vectorization. In brief, Tf-Idf assign weights to the words with respect to their importance, which may overcome the shortcoming of count vectorization. With help of Tf-Idf algorithm, for instance, the word ‘excellent’ would weight 0.054, while the weight of the word ‘table’ would be relatively small, 0.006.

Lastly, previous researches mostly used 3-class classification [1, 3, 6, 8]. But this approach limits many things. For instance, the extremely happiness and admiration expressions like “fantastic course, you the best, excellent work” were placed in same category (positive) with expressions like “good course, fair enough, not bad, you tried sir”, while they represent completely different feelings of the student. In same manner, the negative class contains both of complaint messages “you need to listen more, you missed our assignments, I didn’t understand some topics” and complete disappointment messages like “I hate this course, he needs to be changed, she didn’t even attend to classes, sir you need to retire”. Thus, it is more appropriate to use 5-class classification to get more insight from comments. The 5-class can be listed as Admiration, Positive, Neutral, Negative and Disappointment. So, one of main objectives of this study was to develop a reliable classification model which can address both 3-class and 5-class classifications.

1.2. Purpose of the study

Following are main objectives of this research:

1. To conduct a validation test, compare the correlation coefficient between sentiment analysis scores and scale-rated survey scores.
2. To use a large dataset, more than 52,000 comments, to avoid overfitting the classification model.
3. To develop a state-of-art classification models based on both 3-class and 5-class dataset.
4. To compare the performance of two different vectorization approaches; count vectorization and Tf-Idf (feature weighted approach).

2. Review of Related Literature

There are 2 important aspects in this study that existing literature can help to investigate; how the researchers scored comments in their datasets and what type of machine learning techniques were employed by them. So, every literature was evaluated from two different aspects: formation of classes in dataset and which machine learning techniques were used.

One of early front runners in this field is the study conducted by Pang et al. They tried to explore a new field, classification of movie database comments with respect to their content [11]. They proposed to classify into 2 different categories positive and negative comments. Most of recent researches are influenced by research conducted Pang et al, as they draw a guideline on how to conduct machine learning experiment. They are the ones who proposed to use Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF) and Support Vector Machines (SVM) as alternative tools during classification training.

3-class dataset was mostly used method in researches. Kandhro et al used 3 classes with scores of 0,1 and 2 with respect to their positivity or negativity polarity [3]. Atif used scale-rated scores to classify them into 2 classes [8]. Aung and Myo proposed a dictionary which contained reference words and their ranking, like useful(+2), brilliant(+3) and sleepy(-2) [5]. Lalata et al also used 3-class dataset; positive, negative and neutral [1]. Nelson el al used a dictionary with 5 classes, but converted it 3 class format before experiment to increase accuracy [6].

The methodology proposed by Pang et al was commonly employed by many researchers, with some editions. The effectiveness of NB, RF and SVM algorithms were also discussed and advised for future experiments by Altrabsheh et al. [2].

Lalata et al achieved highest classification accuracy of 90.32% with NB, after comparing results obtained by NB, DT, RF and SVM. In this study they also presented results which shows that using unigrams will lead to highest prediction accuracy, as bi-gram and tri-gram methods resulted with very low accuracy [1]. SVM and NB was used by Nelson et al, but the accuracy was as low as 57.7%. Atif achieved 80% of accuracy, but it was already mentioned that his dataset was relatively small [8].

Some other machine learning techniques (deep learning) were also employed in various researches. In one such study, Kandhro et al used Stochastic Gradient Decent (SGD) and Multilayer Perceptron (MLP) apart from NB and SVM. The highest accuracy was achieved by NB, 83% [3]. Deep learning (Long-Short Term Memory) was employed by the same group of researchers and 90% of validation accuracy was achieved [10]. One similar paper to our methodology was proposed by Oza et al, they used ANN to achieve the prediction accuracy of 87%, but the dataset they used was relatively very small, only 101 comments.

The conducted literature review shows that the commonly used supervised learning algorithms like Random Forest, SVM, Naïve-Bayes were employed during researches, while ANN was also tested as the deep learning method. The performance of all of these classification algorithms were tested during experiment to find out best performer during this research. Gradient boosting was also tested as different approach.

3. Instruments

Almost all of the literature mentioned in Section 2 presented the concise information about commonly used machine learning tools in sentiment analysis, so in this section the presentation of concepts of the classification algorithm like NB, DT, SVM and ANN were skipped, referring to previous studies. A brief information about the accuracy measurements, statistical measurement, sentiment analysis tools, programming language and libraries was given.

3.1. Accuracy measurements

Assuming that the prediction output of model is binary True (T) or False (F) with Positive(P) and Negative(N) actual values, then the results can be represented with 4 different outcomes TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). Hence, 3 major accuracy measurements Precision (PR), Recall (RC) and Accuracy (ACC) can be described by formulas given in (1).

$$PR = \frac{TP}{TP+}, RC = \frac{TP}{TP+}, ACC = \frac{TP+}{TP+TN+FP+F} \tag{1}$$

Cross validation (CV). CV is a model validation technique for assessing how the results of a statistical analysis (model) will be generalized to an independent dataset. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. While training the model, the dataset divided into 5 equal sets; four (4) used for training and the last one to test the accuracy. With every training, validation set changes; thus, all 5 sets completed in 5 rounds [12].

3.2. Statistical measurements

Correlation coefficient (2). It is a measure of the linear relationship between and X and Y. In this paper, the correlation coefficients was used to check how well data of sentiment analysis is correlated to scale-rated scores [13, 14].

$$Corr(X,Y) = \frac{\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n-\bar{x})^2}\sqrt{\sum_{n=1}^{N}(y_n-\bar{y})^2}} \tag{2}$$

P-value. The P-value concept was used to support the significance of the correlation coefficient calculation. In brief, in statistics, the P-value is used to prove that the distribution was not formed by chance, but it has statistical significance. Usually value less than 0.05 is acceptance in reasonable confidence interval.

3.3. Sentiment analysis tools

Tokenization. The comments of the students were split into words, tokens.

Lowercasing. All the words were lower cased, as capital letter and lower-case letter are different inputs during programming.

Stemming. To further facilitate word matching, words in student comments are converted to their root word. For example, “rushing” and “rushed” are converted to “rush”.

Removal of irrelevant content. Punctuation and stop words, which are irrelevant for SA, are removed to improve system response time and effectiveness. Irrelevant comments like “Nnnn”, “DgfgDdsd” or “123dwsd” are removed.

Count vectorization (CountV). This technique assigns value “1” if the particular word presents in the sentence and “0” otherwise

Tf-Idf. Tf-Idf stands for “Term Frequency — Inverse Document Frequency”. This is a technique to quantify a word in documents, the weight to each word is generally computed, which signifies the importance of the word in the document and corpus [16]. To put it in more formal mathematical terms, the Tf-Idf score for the word *t* in the document *d* from the document set *D* is calculated as follows (3):

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3}$$

where$f(t,d) = log(1 + freq(t,d))$ and $idf(t,D) = log\left(\frac{N}{count(d \in D : t \in d)}\right)$.

3.4. Programming tools

Python 3 was used as programming language. NLTK was employed for sentiment analysis process [17]. Scikit Learn library was used during machine learning training [18].

4. Data Collection

Ethics and regulations. All ethics and regulations were followed during data collection. The permission of Nile University of Nigeria was obtained. All the lecture names and respective course codes were encrypted to protect the privacy. Students names were removed, and only comments were used.

Dataset. The dataset was collected from 2018/2019 educational session survey evaluation database. Total number of lecturers was 203, total number of courses was 524, almost 4000 students contributed to dataset with total number of 52,571 comments.

Lecturer’s scale rated scores (LSRS). At the end of educational session students were given a survey where they were asked to evaluate Completion of course, Punctuality, Communication, Fairness, Knowledge of the course on 5-point scale rating survey. The average of 5-point scores were converted to percentage, and each lecturer had his/her own general average over 100. During the experiment, these scores were called Lecturer’s Scale Rated Score (LSRS). LSRS was also extracted from database and was used in correlation check between SA results.

5. Methodology

The following steps were followed during methodology section:

A. Scoring the comments and dataset formation
B. Validity of the sentiment analysis scores
C. Pre-processing of the dataset.
D. K-fold training
E. Grid search for best parameters
F. Testing ANN architectures

5.1. Scoring the comments and dataset formation

All the presented literature in Section 2 were the guideline during scoring the comments [1, 3, 11]. Two different datasets were formed during the study: Dataset 1, where all comments were classified manually into three classes (Positive, Neutral, Negative), and Dataset 2, where the comments were classified manually into five classes (Admiration, Positive, Neutral, Negative and Disappointment). The reason for the formation of the Dataset 1 was to compare the prediction accuracy of newly developed model with existing state-of-art models’ accuracies. Dataset 2 was formed to show that it is still possible to achieve high prediction accuracy, even using 5-class dataset, which was not done before, to the best of our knowledge.

As Pang et al proposed, 2 senior students from English Language major were invited to present guidelines on how to classify the comments with respect to 3-class model and 5-class model.

Dataset 1 (D3). All comments are classified into 3 categories, with respective scores of +1,0,-1. So how big is the integer, so is positive the comment (Positive comment, Neutral comment, Negative comment). Table 1 presents some samples and their scores for D1.

Table 1. Comment samples and their respective scores for D3

COMMENTS	SCORE
Fantastic job, Excellent, You the best sir, My favorite lecturer, He was good, Nice course	+1
No comment, Average lecturer, Manageable course	0
He was late many times, She didn't answer me, Bad one, Poor, Change him, You should retire	-1

Dataset 2 (D5). All comments are classified into 5 categories, with respective scores of +2,+1,0,-1,-2. So how big is the integer, so is positive the comment (Admiration, Positive comment, Neutral, Negative comment and Complete Disappointment). Table 2 presents some samples and their scores.

Table 2. Comment samples and their respective scores for D5

COMMENTS	SCORE
Fantastic job, Excellent, You the best sir, Best ever	+2
He was on good level, He tried, Nice course	+1
No comment, Average lecturer, Manageable course	0
He is always late, She didn't answer me, Poor	-1
Change him, You should retire, Terrible, Worst ever	-2

5.2. Validity of the sentiment analysis scores

To answer to first objective of this research, the LSRS were collected first. Secondly, the average score of comments for the lecturer was calculated for D5, which was shortened as LCSA (Lecturer's Comment Score Average). The correlation and significance test was conducted.

Table 3.Correlation between LSRS and LCSA

		LCSA
LSRS	Pearson Correlation	.799**
	Sig. (2-tailed)	.000
	N	203
**. Correlation is significant at the 0.01 level (2-tailed).		

Result of the Pearson correlation (Table 3) indicated that there was a significant positive association between LSRS and LCSA, (*r(203) = .79, p<.001*). Thus, LCSA can be accepted as reasonable method for course and lecturer evaluation.

5.3. Preprocessing

Before conducting the machine learning experiment, the dataset passed through several pre-processing steps using sentiment analysis tools. Fig. 1 presents used architecture during pre-processing and training. The obtained data passed through pre-processing and each word was tokenized and assigned with weight at the end of the process.

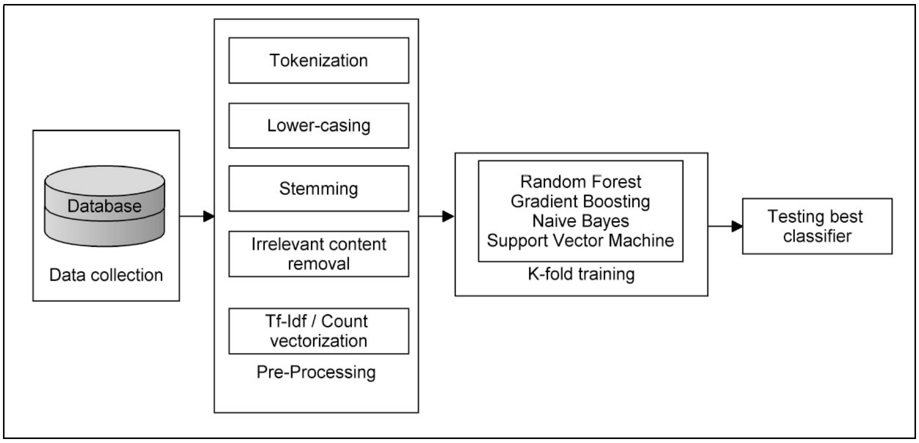

Fig. 1. Architecture of pre-processing and training process.

The dimensions of the dataset after preprocessing was 49287 x 2543 in matrix format. 49287 comments with 2543-word features in each. For instance, assume that the sentence “Sir I am truly concerned about the curriculum of the course” is entering the preprocessing step. Following operations are done to the text (as an example):

Tokenization: [Sir, I, am, truly, concerned, about, the, curriculum, of, the, course]

Lower casing: [sir, i, am, truly, concerned, about, the, curriculum, of, the, course]

Stemming: [sir, i, am, true, concern, about, the, curriculum, of, the, course]

Removal of the stop-words: [sir, i, true, concern, about, curriculum, course]

Vectorization: Count – [0,0,..,1,0,0,..,1,1,0,0…,0,1,0], Tf-Idf – [0,0,..,0.024,0,0,..,0.5,0.006,0,0…,0,0.014,0]

5.4. K-fold training

5-fold cross validation was used to test the accuracy and to find out the best possible classifier. All algorithms were tested with default parameters provided by Scikit Learn [18]. Both of the datasets, D3 and D5, were tested, while both Count Vectorization (CountV) and Tf-Idf vectorization techniques formed two more subsets. Table 4 shows the results for all tested algorithms (rounded to closest integer).

Table 4. Results of initial testing

Model	D5		D3	
	CountV	**Tf-Idf**	**CountV**	**Tf-Idf**
Random Forest	81%	82%	89%	88%
Gradient Boosting	79%	81%	86%	84%
Support Vector Machines	66%	73%	84%	83%
Naïve Bayes	14%	27%	31%	28%

As it stands, RF looks to be better option in comparison to other ones, thus, RF was selected as an algorithm to develop a prediction model. Even though the existing literature shows that NB was one of top performers, the conducted experiments shows that it is not likely to performer well with dataset of big size [1, 3]. Moreover, RF used less time for training comparing to both Gradient boosting and SVM. Next, a grid search method was used to test set of parameters to develop a RF classifier model with highest prediction accuracy.

5.5. Grid search for best parameters (D5)

Several parameters for RF were tested to find the best prediction model. 5-fold cross validation was used to avoid overfitting. 3 different parameters for the number of estimators (10,100, 250) and 4 parameters for the maximum depth of the trees (20, 50, 100, None) were selected, making it total of 12 different sets of parameters.

Table 5 present results of Top 5 models with highest accuracies and their relative parameters for D5 dataset.

Table 5. Results of grid search with RF for D5 dataset.

Models	**Tf- Idf**				**CountV**			
	Max depth	**Number of estim.**	**Mean fit time**	**Mean test score**	**Max depth**	**Number of estim.**	**Mean fit time**	**Mean test score**
Model 1	None	250	560s	90.30%	50	250	609s	85.30%
Model 2	None	100	333s	90.30%	50	100	244s	85%
Model 3	None	10	35s	90.00%	100	250	920s	84.50%
Model 4	100	100	293s	89.60%	100	100	347s	84.50%
Model 5	100	250	732s	89.60%	50	10	28s	83.60%

At this point, the trade between accuracy and training/testing time should be done. For Tf-Idf, even though the Model 1 and 2 overperformed Model 3 with small difference, still Model 3 seems to be the optimal result as the time it used for training and testing is at least 10 times smaller than other two. Thus, the final RF prediction model was built using parameters of Model 3 for Tf-Idf. The same trade off can be assumed for CountV, where Model 2 has optimal training time in comparison to Model 1 and 3, while their test accuracies are almost same.

To test the prediction accuracy of developed model, as a rule of thumb, the dataset was divided in ration of 75% for training the model and 25% for testing. Table 6 presents test accuracy measurements for the developed models.

Table 6. Accuracies for the prediction model with RF for D5.

Technique	**Depth**	**No.Est.**	**Precision**	**Recall**	**Accuracy**
Tf-Idf	None	10	91.70%	91.70%	91.70%
CountV	50	100	90.40%	89.60%	89.60%

5.6. Grid search for best parameters (D3)

The same grid search and testing process was done for D3 dataset and Table 7 presents results of the search.

Table 7. Results of grid search with RF for D3 dataset.

Models	**Tf- Idf**				**CountV**			
	Max depth	**Number of estim.**	**Mean fit time**	**Mean test score**	**Max depth**	**Number of estim.**	**Mean fit time**	**Mean test score**
Model 1	None	100	653s	92.6%	100	100	399s	89.9%
Model 2	None	250	907s	92.5%	100	250	971s	89.8%
Model 3	None	10	63s	92.3%	50	250	726s	89.6%
Model 4	10	250	1408s	92.2%	50	100	247s	89.4%
Model 5	100	100	560s	92%	100	10	40s	89.4%

With the trade-off between accuracy and training time, Model 3 is seeming to be optimal for Tf-Idf, while Model 5 is selected for CountV. Table 8 presents test accuracy measurements for the developed models.

Table 8. Accuracies for the prediction model with RF for D3.

Technique	**Depth**	**No.Est.**	**Precision**	**Recall**	**Accuracy**
Tf-Idf	None	10	96.7%	96.8%	96.8%
CountV	100	10	95.8%	95.9%	95.9%

5.7. Keywords

Fig. 2 presents keywords with highest weights after Tf-Idf processing. Indeed, these words summarize the classes formed during manual classification of the comments. While words like Average, Good, Interest represents positive comments class, Satisfactory and None represent neutral comments class.

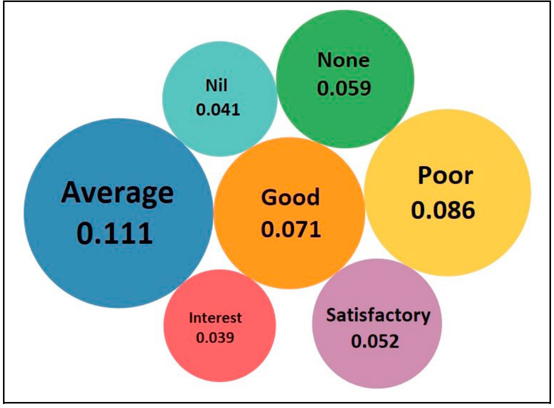

Fig. 2. Keywords and their weights

5.8. Artificial Neural Network

Three different ANN models were tested during experiment. All hidden layers of the network were configured with Rectifier Linear Units (ReLUs). Cross entropy was used as loss function and weights were optimized using Stochastic gradient decent (SGD). Unlike other layers, output layer used Softmax function for final probability prediction [19]. The major difference between them were the number of hidden layers; 1 hidden layer (ANN-1), 2 hidden layers (ANN-2) and 3 hidden layers (ANN-3). Table 9 is the short summary of each ANN architecture.

Table 9. Overview of all ANN architectures

Model	Layer 1	Layer 2	Layer 3	Output
ANN 1	512 nodes	-	-	3/5 classes
ANN 2	2048 nodes	128 nodes	-	3/5 classes
ANN 3	2048 nodes	256 nodes	32 nodes	3/5 classes

The label data was converted to categorical to suit the network. All weights were initiated with uniform distribution. All three models were trained with 200 epochs.

5.9. ANN for D3

All three models successfully decreased the loss value below 0.07, while ANN2 was best performer. All three models reached higher than 98% of validation accuracy, while both ANN2 and ANN3 reached almost same accuracy, 98.9%. Later, all three developed models were tested with unseen data. As a rule of thumb, 75% of the data was used to train the network, while remaining 25% was used to test the prediction accuracy. Table 10 present results for testing experiment for D3.

Table 10. Accuracy measurements for D3 testing data

Model	Precision		Recall		Accuracy	
	Tf-Idf	CountV	Tf-Idf	CountV	Tf-Idf	CountV
ANN1	96%	97%	96%	96%	96%	97%
ANN2	97%	97%	97%	97%	97%	97%
ANN3	96%	97%	96%	96%	96%	97%

5.10. ANN for D5

Next, the D5 dataset was tested with ANN. The developments for both loss function and validation accuracy were almost same with D3, presenting very promising results. Table 11 presents comparative results of loss function and validation accuracy on both datasets.

Table 11. Final loss values and validation accuracies for both D3 and D5 dataset

	Dataset	**ANN1**	**ANN2**	**ANN3**
Loss value	D3	0.0671	0.0432	0.0548
	D5	0.22	0.1841	0.192
Validation Accuracy	D3	97.90%	98.80%	98.80%
	D5	93.40%	95%	95%

Later, all three developed models were tested with unseen data just like with D3 dataset. Table 12 present results for testing experiment for D5 dataset.

Table 12. Accuracy measurements for D5 testing data

Model	**Precision**		**Recall**		**Accuracy**	
	Tf-Idf	**CountV**	**Tf-Idf**	**CountV**	**Tf-Idf**	**CountV**
ANN1	92%	92%	91%	92%	92%	92%
ANN2	93%	93%	92%	92%	92%	92%
ANN3	93%	93%	92%	92%	92%	92%

For both datasets, D3 and D5, the model ANN2 performed on same level with ANN3, while using less training time, as the structure of ANN3 is deeper. Hence, to avoid the computationally expensive model, ANN2 was selected as proposed model.

6. Discussion

The summary of all experiments with respective models presented on Table 13. The presented values are the results of prediction accuracies with the best parameters for each vectorization technique. Prediction accuracies of RF models were rounded to closest integer. The results presented on Table 13 shows that RF and ANN performed almost on same level. In this case, the preference may be given to RF, as it is simpler model and it is not computationally expensive technique like ANN. Still ANN remains as very reliable tool with highest prediction accuracy. The measures of precision, recall and accuracy were on balanced level for both of RF and ANN models, which shows that prevention of development of overfitted model was successfully done.

Table 13. Summary of the experiment

Model	**D3**		**D5**	
	Tf-Idf	**CountV**	**Tf-Idf**	**CountV**
Random Forest	97%	96%	92%	90%
ANN	97%	97%	92%	92%

On the other hand, while Tf-Idf outperformed CountV during classification model development with Random Forest, no such effect was observed during ANN experiment. This is definitely explaining the superiority of ANN to adopt the input values by changing the weights between nodes. No such feature is presents in Random Forest, as it uses input numbers straight. Consequently, Tf-Idf can be proposed as effective vectorization method for student-lecturer comment classification.

Table 13 results for D3 dataset appears to be state-of-art results, as the prediction accuracy of 97% is higher than any previous research results. Moreover, a reliable model with prediction accuracy of 92% was developed for D5 dataset, which supports one of main objectives of this study, to develop an efficient 5-class classification model.

7. Conclusion

In this paper, the comparative study was conducted on different vectorization and classification methods to built state-of-art classification model to classify lecturer-student comments. Tf-Idf and Count vectorization were compared as text-to-vector methods, while Random forest, Support vector machine, Naïve- Bayes, Gradient boosting and Artificial neural networks were tested as classification algorithms. The found results shows that the combination of RF with Tf-Idf resulted with one of highest prediction accuracies with optimal training time, while ANN slightly overperformed RF model and selection of vectorization method didn't affect the results too much during deep learning training. The prediction accuracy for 3-class dataset was as high as 97%, while the accuracy for 5-class dataset reached very promising 92% level.

Further researches are needed in the future, as there are still some gaps need to be addressed. Unigram method was tested during this experiment, so bigrams or trigrams can be also tested, even though Lalata et stated that usage of bigrams or trigrams during sentiment analysis affected the accuracy in negative way [1].

References

[1] Lalata JP, Gerardo B, Medina R. A Sentiment Analysis Model for Faculty Comment Evaluation Using Ensemble Machine Learning Algorithms. In: *Proceedings of the 2019 International Conference on Big Data Engineering*. 2019, pp. 68–73.

[2] Altrabsheh N, Gaber MM, Cocea M. SA-E: sentiment analysis for education. In: *International conference on intelligent decision technologies*. 2013, pp. 353–362.

[3] Kandhro IA, Chhajro MA, Kumar K, et al. Student Feedback Sentiment Analysis Model using Various Machine Learning Schemes: A Review. *Indian Journal of Science and Technology* 2019; **12(14)**.

[4] Rani S, Kumar P. A sentiment analysis system to improve teaching and learning. *Computer* 2017; **50**: 36–43.

[5] Aung KZ, Myo NN. Sentiment analysis of students' comment using lexicon based approach. In: *2017 IEEE/ACIS 16th international conference on computer and information science (ICIS)*. IEEE, 2017, pp. 149–154.

[6] Cunningham-Nelson S, Baktashmotlagh M, Boles W. Linking numerical scores with sentiment analysis of students' teaching and subject evaluation surveys: Pointers to teaching enhancements. In: *27th Annual Conference of the Australasian Association for Engineering Education: AAEE 2016*. Southern Cross University, 2016, p. 187.

[7] Li C, Ma J. Research on online education teacher evaluation model based on opinion mining. In: *2012 National Conference on Information Technology and Computer Science*. Atlantis Press, 2012.

[8] Atif M. An Enhanced Framework for Sentiment Analysis of Students" Surveys: Arab Open University Business Program Courses Case Study. *Business and Economics Journal* 2018; **9**: 337.

[9] Oza KS, Kamat RK, Naik PG. Student feedback analysis: a neural network approach. In: *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, 2017, pp. 342–348.

[10] Kandhro IA, Wasi S, Kumar K, et al. Sentiment analysis of students' comment using long-short term model. *Indian J Sci Technol* 2019; **12**: 1–16.

[11] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume* **10**. Association for Computational Linguistics, 2002, pp. 79–86.

[12] Ochilbek R. Using data mining techniques to predict and detect important features for book borrowing rate in academic libraries. In: *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*. IEEE, 2019, pp. 1–5.

[13] Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. Springer series in statistics New York, 2001.

[14] Ochilbek R. Development of a Method for Evaluating Quality of Education in Secondary Schools Using ML Algorithms. In: *Proceedings of the 2019 11th International Conference on Education Technology and Computers*. 2019, pp. 23–29.

[15] Dahiru T. P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine* 2008; **6**: 21–26.

[16] Ramos J. Using tf-idf to determine word relevance in document queries. In: *Proceedings of the first instructional conference on machine learning*. Piscataway, NJ, 2003, pp. 133–142.

[17] Loper E, Bird S. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

[18] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 2011; **12**: 2825–2830.

[19] Gulli A, Pal S. *Deep Learning with Keras*. Packt Publishing Ltd, 2017.