Sentiment Analysis of Students' Feedback on Faculty Online Teaching Performance Using Machine Learning Techniques

CAREN AMBAT PACOL
Information Technology Department,
Pangasinan State University,
San Vicente, Urdaneta City, Pangasinan,
PHILIPPINES

Abstract: - The pandemic has given rise to challenges across different sectors, particularly in educational institutions. The mode of instruction has shifted from in-person to flexible learning, leading to increased stress and concerns for key stakeholders such as teachers, parents, and students. The ongoing spread of diseases has made in-person classes unfeasible. Even if limited face to face classes will be allowed, online teaching is deemed to remain a practice to support instructional delivery to students. Therefore, it is essential to understand the challenges and issues encountered in online teaching, particularly from the perspective of students. This knowledge is crucial for supervisors and administrators, as it provides insights to aid in planning intervention measures. These interventions can support teachers in enhancing their online teaching performance for the benefit of their students. A process that can be applied to achieve this goal is sentiment analysis. In the field of education, one of the applications of sentiment analysis is in the evaluation of faculty teaching performance. It has been a practice in educational institutions to periodically assess their teachers' performance. However, it has not been easy to take into account the students' comments due to the lack of methods for automated text analytics. In line with this, techniques in sentiment analysis are presented in this study. Base models such as Naïve Bayes, Support Vector Machines, Logistic Regression, and Random Forest were explored in experiments and compared to a combination of the four called ensemble. Outcomes indicate that the ensemble of the four outperformed the base models. The utilization of Ngram vectorization in conjunction with ensemble techniques resulted in the highest F1 score compared to Count and TF-IDF methods. Additionally, this approach achieved the highest Cohen's Kappa and Matthews Correlation Coefficient (MCC), along with the lowest Cross-entropy, signifying its preference as the model of choice for sentiment classification. When applied in conjunction with an ensemble, Count vectorization yielded the highest Cohen's Kappa and Matthews Correlation Coefficient (MCC) and the lowest Cross-entropy loss in topic classification. Visualization techniques revealed that 65.4% of student responses were positively classified, while 25.5% were negatively classified. Meanwhile, predictions indicated that 47% of student responses were related to instructional design/delivery, 45.3% described the personality/behavior of teachers, 3.4% focused on the use of technology, 2.9% on content, and 1.5% on student assessment.

Key-Words: - Machine learning, max voting ensemble, natural language processing, online teaching, sentiment analysis, teaching performance, vectorization, visualization

Received: April 27, 2023. Revised: November 26, 2023. Accepted: December 23, 2023. Published: February 19, 2024.

1 Introduction

The onset of the COVID-19 pandemic caused a substantial disruption in higher education as institutions were compelled to shift to online learning, mandated by lockdowns. Despite the gradual improvement in the epidemiological situation, online learning continues to gain popularity, offering novel educational opportunities, [1]. Therefore, it is essential to understand the challenges and issues encountered in online teaching, particularly from the perspective of students. This knowledge is crucial for supervisors and administrators, as it provides insights to aid in planning intervention measures. These interventions can support teachers in enhancing their online teaching performance for the benefit of their students. A process that can be applied to achieve this goal is sentiment analysis.

Sentiment Analysis has been a very interesting area in research since it reveals opportunities to learn and improve customer experiences and build better products. This motivated practitioners and researchers in various areas to apply sentiment analysis. In the field of education, one of the

applications of sentiment analysis is the evaluation of the quality of instruction, [2]. Monitoring the perspective of students on the quality of instruction provided by their teachers is very significant for all role-players including teachers, students, and administration. It has been a practice for educational institutions to evaluate their teachers' performance periodically. However, it has not been easy to take into account the students' comments due to the lack of methods for automated text analytics. Analyzing the textual feedback of students and interpreting them whether positive or negative is important since it provide insights on the satisfaction of student/s on teaching performance when summarized. It is also worthwhile to understand the aspect of teaching that is described in student responses as this will aid immediate supervisors in pinpointing where the teacher needs improvement. Additionally, it gives insights into specific concerns that bother most students when it comes to online teaching/learning. The aspects of teaching that were considered in this study are identified in section 2.2. Textual feedback from a larger population of students is preferred to get reliable results. However, as the number of textual data increases, manually analyzing them also becomes tedious. This calls for the utilization of sentiment analysis.

Researchers have performed sentiment analysis mostly to classify sentiments and describe the viewpoints of students regarding the teaching performance of their teachers. In [3], the authors employed a qualitative methodology to identify the most commonly used words in describing the teaching performance of educators in online courses. The authors in [4], presented the most frequently occurring words related to student sentiments, along with the emerging clusters generated from those sentiments. In [5], they carried out sentiment analysis using Knime and found positive sentiments demonstrated superior performance, achieving the highest recall and precision rates. Previous studies proved beneficial as they provided insights on conducting sentiment analysis for teaching performance. However, they were constrained by certain limitations. In [3], [4], and [5], sentiment analyses were conducted classifying sentiments as positive, negative, or neutral. The authors did not include an analysis based on the various aspects of teaching. In [3] and [5], their visualization of results was based on unigrams only. Visualization of sentiments based on bigrams and higher order Ngrams provides a more nuanced understanding of sentiment by considering pairs or groups of words. This helps capture the context in which words are used, enhancing the accuracy of sentiment analysis. In this paper, aspects of teaching were integrated into the sentiment analysis and visualization of sentiments using bigrams and trigrams instead of unigrams are presented.

Two common approaches to sentiment analysis are lexicon-based and machine learning-based techniques. The lexicon-based approach utilizes a dictionary of sentiment words that are assigned pre-defined weight to specify its sentiment polarity. On the other hand, machine learning focuses on producing algorithms that can be used in artificial intelligence applications in the actual world. The increasing size of data in enterprises caused the necessity of using machine learning techniques to discover intelligence in business which aids in strategic decision-making, [6]. The advantage of machine learning is it provides the ability to train the algorithms. The availability of sentiment packages along with sentiment corpora and various manually annotated sentiment rules combined with Natural Language Processing (NLP) opens the opportunity to produce improved, faster, and more accurate algorithms. Though machine learning methods are great, they have limitations and are not capable of working at a character level like humans. Transformations are necessary on the original data so that it can be processed more easily by the machine learning method, thereby improving the performance of the methods.

Several studies have been conducted performing machine learning-based sentiment analysis. A study by [7] used nine machine learning algorithms in their experiments applying bag-of-words and TF-IDF vectorization. They found that algorithms in use do not yet precisely classify neutral sentiments and concluded that more datasets containing educational content are required to enhance sentiment analysis algorithms. In [8], the authors used support vector machines (SVM) with Ngram and TF-IDF vectorization in sentiment analysis to investigate students' perceptions of e-Learning. They concluded that the SVM classifier successfully predicted positive and negative sentiments in tweets, demonstrating an overall commendable performance. Another study by [9], employed the majority voting principle integrating Naive Bayes, Logistic Regression, Support Vector Machines, Decision Tree, and Random Forest algorithms while utilizing Ngram analysis and the TF-IDF method. Their findings indicated that the ensemble classification system outperformed the individual classifiers. Although studies on evaluating the performance of machine learning algorithms have been conducted, research on evaluating the

performance of an ensemble of machine learning algorithms applied with various vectorization techniques in teaching and learning is still limited. Through this study, the author hopes to contribute new insights to the field by employing machine learning techniques in the analysis of teaching performance sentiments.

The objectives of this study are: (a) assess the performance of the base models as compared to the ensemble model in students' textual feedback classification (b) explore and assess the performance of TF-IDF and Ngram techniques when used in base models and ensemble techniques and (c) provide visualization of students' sentiments in the online teaching performance evaluation.

2 Methodology

The methodology is divided into the following steps.

2.1 Data Gathering

Online teaching performance data of 109 faculty members during the first semester and 129 faculty members during the second semester of the previous academic year were sourced from a campus of a state university in the Philippines.

2.2 Data Preparation

Comments were taken from data gathered and summarized in an Excel file. For the first annotation, cleaning of data was done manually wherein misspelled words were corrected. Manual labeling was performed on a total of 18,004 sentences to generate the training dataset. Sentences were annotated using marks such as 1 for positive, -1 for negative, and 0 for neutral. The training dataset consists of 11483 positive sentences, 4781 negative sentences, and 1740 neutral sentences indicating that the dataset was unbalanced. For the second annotation, a total of 16485 sentences were left after cleaning. Some sentences cannot be classified in any of the indicators and as such were removed from the dataset. Aspects of teaching performance that were used to label sentences for topic classification were identified based on Checklist for Evaluating Online Courses, [10]. The sentences were labeled with content (C), instructional design/delivery (ID), student assessment (SA), use of technology (T), and personality/behavior (PB). Personality/Behavior was included as one of the aspects of teaching performance during the annotation as it was observed some sentences pertains to the character and attitude of teachers. The dataset consists of 770 content-related comments, 7394 instructional design/delivery related sentences, 7224 comments related to faculty personality/behavior, 703 sentences on the use of technology, and 394 comments on students' assessment.

2.3 Data Pre-processing

After data preparation, the Natural Language Toolkit (NLTK) package in Python were used for pre-processing. Sentences were pre-processed by removing special characters, substituting multiple spaces with single spaces, removing prefixes, and conversion of all text to lowercase. Stop words were also removed.

2.4 Training the Model

The pre-processing was followed by training the model to perform classification using the labeled dataset. The machine learning algorithms used as base models were naïve bayes, support vector machines, logistic regression, and random forest.

2.5 Testing the Model and Measuring Performance

Utilizing 25% of the data for testing, the trained model was evaluated yielding the accuracy, precision, recall, and F1 score. A function called classification_report() of sklearn.metrics in python was used.

2.6 Applying TF-IDF, Ngrams and Ensemble Machine Learning Techniques in Sentiment Classification

An experiment was conducted applying TF-IDF, Ngrams, and ensemble techniques to improve the classifier. Ngrams and TF-IDF are text vectorization approaches. The process of vectorization converts text into a form that the machine can understand. The base models were applied with a Count/bag-of-words vector. In Count vectorization, words and the number of their occurrences in a document are generated, [11]. Separate experiments were also observed applying TF-IDF and Ngrams.

Every sentence is represented as binary vectors in Count vectorization. Meanwhile, more information is encoded into the vector using TF-IDF. Term Frequency-Inverse Document Frequency (TF-IDF) assesses the importance of a word within a corpus of textual data. Term Frequency (TF) measures the similarity among documents, while Inverse Document Frequency (IDF) gauges the significance of a word. Thus, the product of TF and IDF for a word yields the frequency of that word in the document multiplied by the uniqueness of the

word, [12]. On the other hand, the Ngrams refer to collections of words grouped by 1 in the case of unigrams, 2 for bigrams, 3 for trigrams, and so forth. For instance, the sentence "You are good" is transformed into a vector (“You”, “are”, “good”) in unigrams and (“you are”, “are good”) in bigrams, [13]. Ngram range of 1 and 2 were set in an experiment converting sentences to unigram and bigram vectors.

An experiment on combining single models in one model using Max Voting Ensemble was also conducted. In the Max Voting Ensemble, various models were used to predict classification. The prediction of every model is called a ‘vote’. The final prediction is taken based on the predictions from the majority of the models, [14]. A classification report is then generated and its performance is compared to the base models. Ngram and TF-IDF were combined with the ensemble in two separate experiments. Figure 1 illustrates the necessary input, processes to execute, and the anticipated output in the ensemble model.

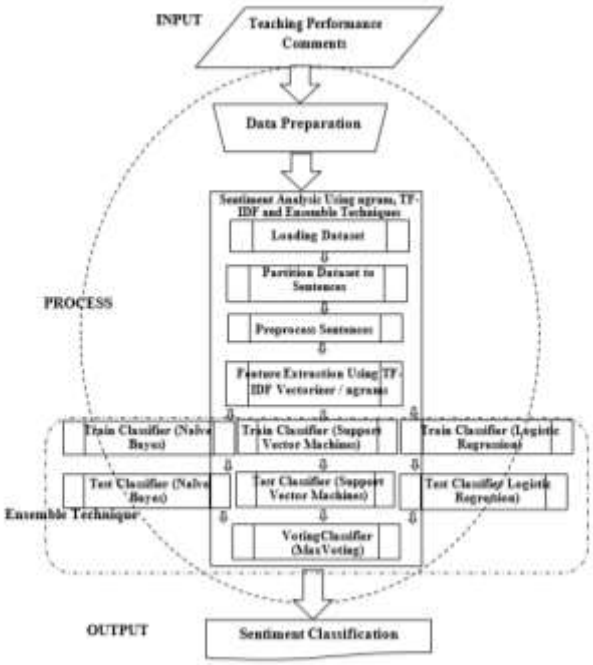

Fig. 1: Sentiment Classification using Ngram, TF-IDF and Ensemble Techniques

2.7 Evaluating the Classification Model

A confusion matrix and a classification report were generated to evaluate the performance of the classification model. The dataset was unbalanced both for sentiment classification and topic classification, thus, weighted precision, recall, and F1 score from the classification report were used in the evaluation. Cohen’s Kappa, Cross-entropy loss, and Matthew’s Correlation Coefficient (MCC) were also calculated to further support the evaluation results. All of these are metrics that can be used for evaluating multi-class classifiers on unbalanced datasets.

Cohen's Kappa calculates a score that indicates the degree of agreement between two annotators, [15]. The Cross-entropy loss used in (multinomial) Logistic Regression is defined as the negative log-likelihood of a logistic model that returns y_pred probabilities for its training data y_true. The y_true represents the correct labels for the samples while the y_pred are the predicted probabilities, [16]. The Matthews Correlation Coefficient considers both true and false positives and negatives. A coefficient of +1 signifies a perfect prediction, 0 indicates an average random prediction, and -1 represents an inverse prediction, [17].

2.8 Visualization of Students’ Sentiments in the Online Teaching Performance Evaluation

Visualization techniques such as bar graph, pie chart, and word cloud were used in Jupyter Notebook importing matplotlib.pyplot and wordcloud. Frequent phrases from positive and negative comments were extracted and presented through bar graphs and are further supported by wordcloud of bigrams and trigrams. Overall sentiments expressing percentages of positive, negative, and neutral sentences were shown in a pie chart.

3 Results and Discussion

This section presents and deliberates on the outcomes uncovered in this study. The results are divided into two sections. First, the results of evaluation on base and ensemble models in sentiment classification. Second, the results of evaluation on base and ensemble models in topic classification. In addition, visualization of sentiments in both sentiment and topic classification was also presented.

3.1 Results of Evaluation of Base and Ensemble Models in Sentiment Classification

The performance of the machine learning algorithms was observed when Count (bag-of-words), TF-IDF, and Ngram vectorization techniques were applied in the dataset and after pre-processing steps were performed. Accuracy, macro, and weighted averages of scores for precision, recall, and F1 were calculated in the classification report. Macro average computes metrics individually for each label and

then determines their unweighted mean across all labels. This does not take label imbalance into account. Meanwhile, weighted averages calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance and it can result in an F-score that is not between precision and recall. The efficacy of classification accuracy is optimal when the number of samples is evenly distributed across each class. Since there is an imbalance in the number of instances in each class, the weighted average results for precision, recall, and F1 score in the training dataset applying Count (bag-of-words), TF-IDF, and Ngrams were considered.

Higher weighted average precision, recall, and F1 score were desired in this case. Precision is calculated as the ratio of True Positives to the sum of True Positives and False Positives. This means precision is a measure of the classifier's exactness. This shows a low precision and indicates a large number of False Positives. On the other hand, recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. This means recall is a measure of the classifier's completeness. This shows a low recall indicates many False Negatives. Meanwhile, F1 Score is the 2*((precision * recall)/(precision + recall)). The F1 score reflects a harmonious balance between precision and recall.

Results of Count vectorization on sentiment classification illustrate that among the base models, higher weighted average precision, recall, and F1-score on both Logistic Regression and Support Vector Machines were obtained. However, the ensemble outperformed Logistic Regression and Support Vector Machines yielding 0.87 precision, 0.88 recall, and 0.87 F1 scores compared to Logistic Regression and Support Vector Machines that obtained the same precision, recall, and F1 scores of 0.86.

Results on TF-IDF vectorization show that Support Vector Machines yielded the highest precision and recall of 0.87 but the same F1 score of 0.86 in the two base models, Support Vector Machines and Logistic Regression. Logistic Regression obtained precision and recall of 0.86. Naïve Bayes yielded a precision of 0.85, recall of 0.84, and F1 score of 0.83. Random Forest got precision, recall, and F1 scores of 0.85. Meanwhile, the ensemble obtained 0.86 in precision, recall, and F1 scores. This indicates that the ensemble did not outperform Support Vector Machines and Logistic Regression in terms of F1 score. TF-IDF increased the precision and recall of Support Vector Machines by 0.01 but no improvement was found in the F1 score. It also did not improve the precision, recall, and F1 scores of the other machine learning algorithms including the ensemble. A study by [18], found that machine learning models generally achieved higher accuracy rates with TF-IDF, except in the cases of Multinomial Naïve Bayes and Neural Network I, where the Count vectorizer demonstrated superior performance in terms of accuracy percentages. More specifically, within these two models, TF-IDF exhibits superior performance on the IMDb movie reviews test set, stemming from the dataset on which the models were trained while showing inferior results on other datasets. The other datasets are reviews on clothing, food, hotels, Amazon products, and tweets. In another study by [19], TF-IDF demonstrated greater efficiency compared to Count vectorizer when dealing with large-volume datasets. Both vectorizers exhibited approximately similar performance, except for Single Layer Perceptron, where the Count vectorizer achieved a 10% higher accuracy. The findings in this present study are aligned with the findings of [18] and [19] that though TF-IDF vectorizer is often regarded as better than Count vectorizer, it does not generalize in all cases. It is interesting to note that in the two prior studies, TF-IDF and Count vectorizer were applied in various datasets. This suggests that the differences in the performance of the two vectorizers can be attributed to the characteristics of the datasets. For instance, a Count vectorizer might be more effective when the data is shorter and it has fewer distinct words, [20].

Results on Ngram vectorization indicate that Ngrams outperformed Count in terms of F1 score when applied in the ensemble. These findings support the findings of [9], that ensemble yielded better performance in sentiment classification of students' comments than individual classifiers. They used Ngram analysis for feature extraction. The findings in this present study also complement that of [21]. They found that the ensemble model demonstrated a good ability to cope with errors.

The weighted average results in the training dataset applying Ngrams are presented in Table 1. Table 1 illustrates that Ngrams setting ngram_range to 1, 2 (unigrams + bigrams) yielded the highest precision, recall and F1 score when applied in ensemble in the training dataset. A comparison of the F1 score of the ensemble applying Count and Ngrams is shown in Table 2. Results indicate that Ngrams outperformed count in terms of F1 score when applied in the ensemble.

Table 1. Performance of Machine Learning Algorithms Using Ngrams Vectorization in Sentiment Classification

Metric	**Machine Learning Algorithms**				
	Logistic Regression (LR)	**Naïve Bayes (NB)**	**Support Vector Machines (SVM)**	**Random Forest (RF)**	**Ensemble (LR+NB+SVM+RF)**
Accuracy	0.87	0.82	0.87	0.84	0.88
Weighted average					
Precision	0.87	0.82	0.86	0.85	0.88
Recall	0.87	0.82	0.87	0.84	0.88
F1	0.87	0.80	0.86	0.84	**0.88**

Table 2. Summary of Machine Learning Algorithms with Highest F1 Score in Sentiment Classification

Metric	**Machine Learning Algorithms**	
	Ensemble (LR+NB+SVM+RF) + Count	**Ensemble (LR+NB+SVM+RF) + ngrams (1, 2)**
F1	0.87	**0.88**

Predicting positive, negative, and neutral sentences correctly is necessary to provide correct information in the visualization of students' sentiments. Higher recall in all classes is desired since identifying true positives in each class is crucial. Higher precision is also important as it demonstrates the confidence that all predicted positives in each class are true positives. Since both recall and precision are important metrics in this case, the F1 score can be used to convey the balance between them.

Cohen's Kappa, Cross-Entropy Loss, and Matthews Correlation Coefficient (MCC) were utilized to further evaluate the classifiers on the unbalanced dataset. These three are considered more robust statistical metrics, particularly in scenarios with imbalanced class distribution. High scores in both MCC and Cohen's Kappa are achieved when predictions demonstrate favorable outcomes across all four parameters of the confusion matrix (true positives, false negatives, true negatives, and false positives), proportionate to the sizes of positive and negative elements in the dataset. The efficacy of MCC is evident in various scientific journals, including its application in medical diagnostics, [22]. Cross-entropy loss is a commonly employed loss function in classification tasks, assessing the alignment between predicted probabilities and actual probabilities. It gauges the difference between two probability distributions, usually the actual distribution and a predicted or estimated one. A lower loss signifies enhanced model performance, and a Cross-entropy loss of 0 signifies perfection, [23].

Cohen's Kappa was calculated using sklearn.metrics.cohen_kappa_score class setting y1 and y2 parameters to actual and predicted classes respectively. Cross-entropy loss was computed utilizing sklearn.metrics. log_loss calculating first the predicted probabilities using predict_proba method.

The parameters y_true and y_pred were set to actual class and results of predict_proba method respectively. MCC was determined using sklearn.metrics.matthews_corrcoef setting parameters y_true and y_pred to actual and predicted classes respectively.

Ngram vectorization set to ngram_range of 1, 2 (unigrams + bigrams) applied in ensemble yielded Cohen's Kappa and MCC score closest to 1. The Cohen's Kappa of 0.76 indicates that there is substantial agreement between the predicted and actual values. Ensemble also yielded the lowest Cross-entropy loss of 0.36 supporting the findings that ensemble applied with Ngrams is preferred in this case.

Table 3 provides a comparison of these metrics on Logistic Regression, Support Vector Machines, and ensemble as they were able to yield higher results compared to Naïve Bayes and Random Forest based on classification report.

Table 3 presents that Ngrams vectorization set to ngram_range of 1, 2 (unigrams + bigrams) applied in ensemble yielded Cohen's Kappa and MCC score closest to 1. The Cohen's Kappa of 0.76 indicates that there is substantial agreement between the predicted and actual values. The ensemble also yielded the lowest Cross-entropy loss of 0.36. This further supports the findings that an ensemble applied with Ngrams is preferred in this case.

Table 3. Comparison of Cohen's Kappa, Cross-Entropy, and Matthews Correlation Coefficient (MCC)on Logistic Regression, Support Vector Machines, and Ensemble using Count and Ngrams in Sentiment Classification

Machine Learning Algorithm + Text Vectorization Technique	**Metrics Cohen's Kappa**	**Cross-Entropy Loss**	**MCC**
Logistic Regression +Count	0.7220	0.3786	0.7244
Logistic Regression +Ngrams(1, 2)	0.7386	0.3601	0.7420
Logistic Regression +Ngrams(1, 3)	0.7362	0.3642	0.7402
Logistic Regression +Ngrams(1, 4)	0.7337	0.3678	0.7379
Support Vector Machines +Count	0.7299	0.3958	0.7312
Support Vector Machines +Ngrams(1, 2)	0.7331	0.3817	0.7312
Ensemble (Count)	0.7518	0.3641	0.7536
Ensemble ngrams(1, 2)	**0.7586**	**0.3556**	**0.7605**
SVM(tf-idf)	0.7317	0.3639	0.7359

Figure 2 shows that the comments of students on online teaching performance are dominated by positive sentences with 65.4%. However, the 25.5% negative sentences should not also be ignored and have to be addressed to improve learning experiences among students.

Figure 3 demonstrates 3-5 ngrams (trigrams+4-grams+5-grams) representing the most frequent phrases in positive students' comments. Meanwhile, Figure 4 reveals the 3-5 ngrams most frequent phrases in negative students' comments. The top most frequent phrases in positive comments include "*ability to explain difficult things in a simple way*", "*explains subject matter*", "*good in explaining topic*", "*explain the topic well*" and "*magaling magturo*". Top most frequent phrases in negative comments are "*poor internet connection*", "*weakness internet connection*", "*slow internet connection sometimes*", "*weakness time management*", "*lack of time in discussing*", "*sometimes the discussion is fast*" and "*unable to meet us*".

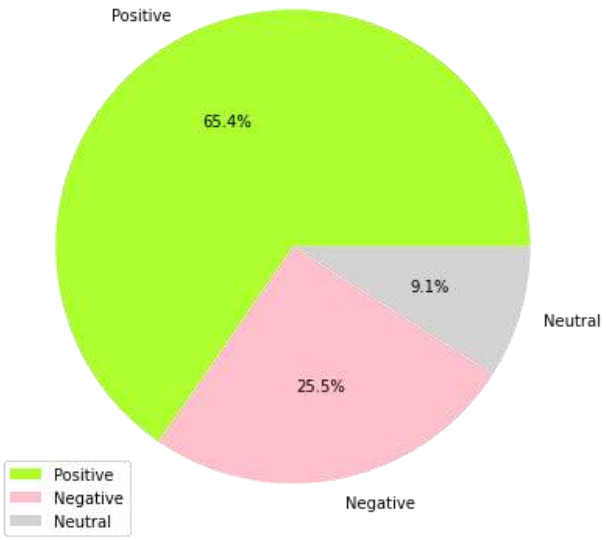

Fig. 2: Overall Sentiments of Students on Online Teaching Performance based on Sentiment Classification

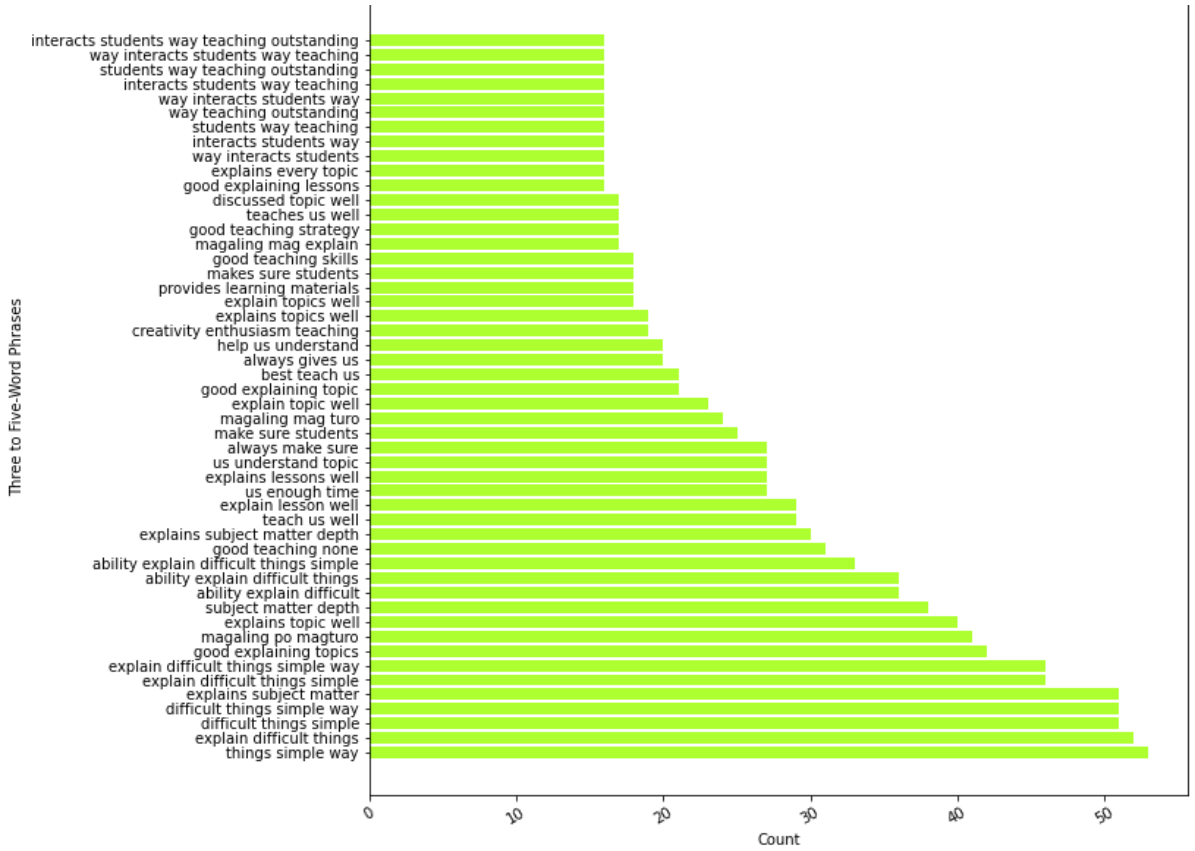

Fig. 3: Top 50 Most Frequent Phrases from Positive Students' Comments

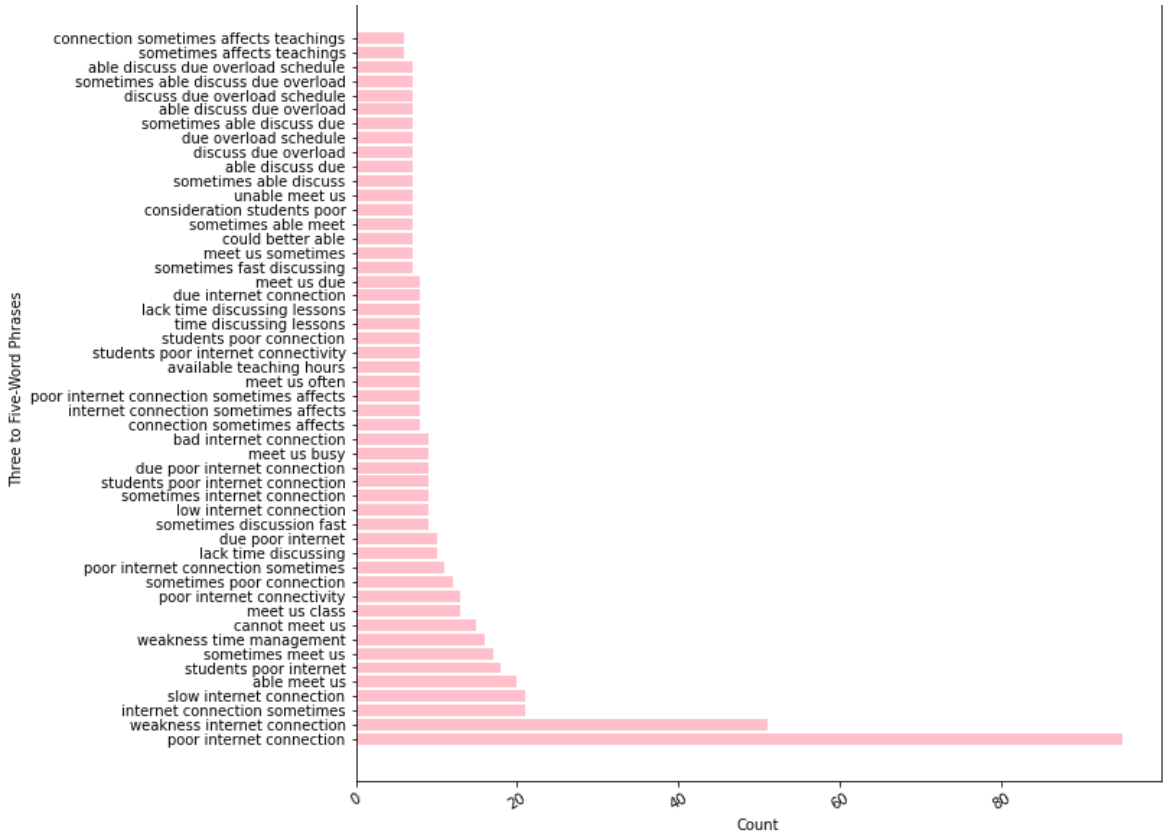

Fig. 4: Top 50 Most Frequent Phrases from Negative Students' Comments

3.2 Results of Evaluation on Base and Ensemble Models in Topic Classification

Similar to the case in sentiment classification, there is an imbalance in the number of instances in each class after labeling them with marks for content, instructional design/delivery, use of technology, student assessment, and personality/behavior. Thus, the weighted average results for precision, recall, and F1 score in the training dataset were considered.

Results show that Ngrams setting ngram_range to 1, 2 (unigrams + bigrams) yielded the highest precision, recall and F1 score when applied in the ensemble in the training dataset. A comparison of the F1 score of the ensemble applying Count and Ngrams is shown in Table 4. Results indicate that the same F1 score was yielded with both Ngrams and Count when applied in the ensemble.

Cohen's Kappa, Cross-entropy loss, and Matthews Correlation Coefficient (MCC) were again utilized to evaluate further the classifiers on the unbalanced dataset. Table 5 provides a comparison of Cohen's Kappa, Cross-entropy, and Matthews Correlation Coefficient (MCC) on Logistic Regression, Support Vector Machines, and ensemble as they were able to yield higher results compared to Naïve Bayes and Random Forest based on the classification report.

Table 4. Summary of Machine Learning Algorithms with Highest F1 Score in Topic Classification

Metric	**Machine Learning Algorithms**	
	Ensemble (LR+NB+SVM+RF) + Count	**Ensemble (LR+NB+SVM+RF) + ngrams (1, 2)**
F1	0.82	0.82

Table 5 indicates that Count vectorization applied in the ensemble resulted in Cohen's Kappa and MCC scores closer to 1 when compared to those of Ngrams. It also got lower Cross-entropy loss than Ngrams.

In Figure 5, it is evident that 47% of the students' responses were predicted to discuss instructional design/delivery, 45.3% delved into the personality/behavior of teachers, 3.4% centered on the use of technology, 2.9% on content, and 1.5% on student assessment.

Bar charts and word cloud were again used to provide visualization of the most frequent phrases in each topic. Examples were provided in Figure 6 and Figure 7, demonstrating 3-5 Ngrams (trigrams+4-grams+5-grams) in instructional design/delivery and student assessment respectively. The actual classification of the sentences was used.

Table 5. Comparison of Cohen's Kappa, Cross-Entropy, and Matthews Correlation Coefficient (MCC) on Ensemble using Count and Ngrams in Topic Classification

Machine Learning Algorithm + Text Vectorization Technique	**Metrics**		
	Cohen's Kappa	**Cross-Entropy Loss**	**MCC**
Ensemble (Count)	**0.7065**	**4.1726**	**0.7076**
Ensemble ngrams(1, 2)	0.6978	5.9045	0.6990

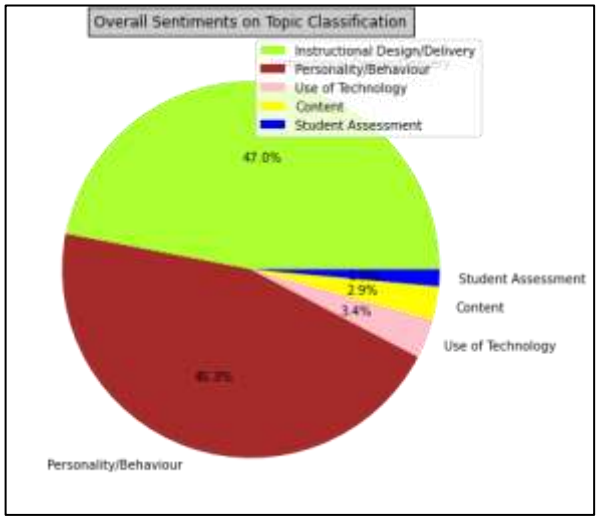

Fig. 5: Overall Sentiments of Students on Online Teaching Performance based on Topic Classification

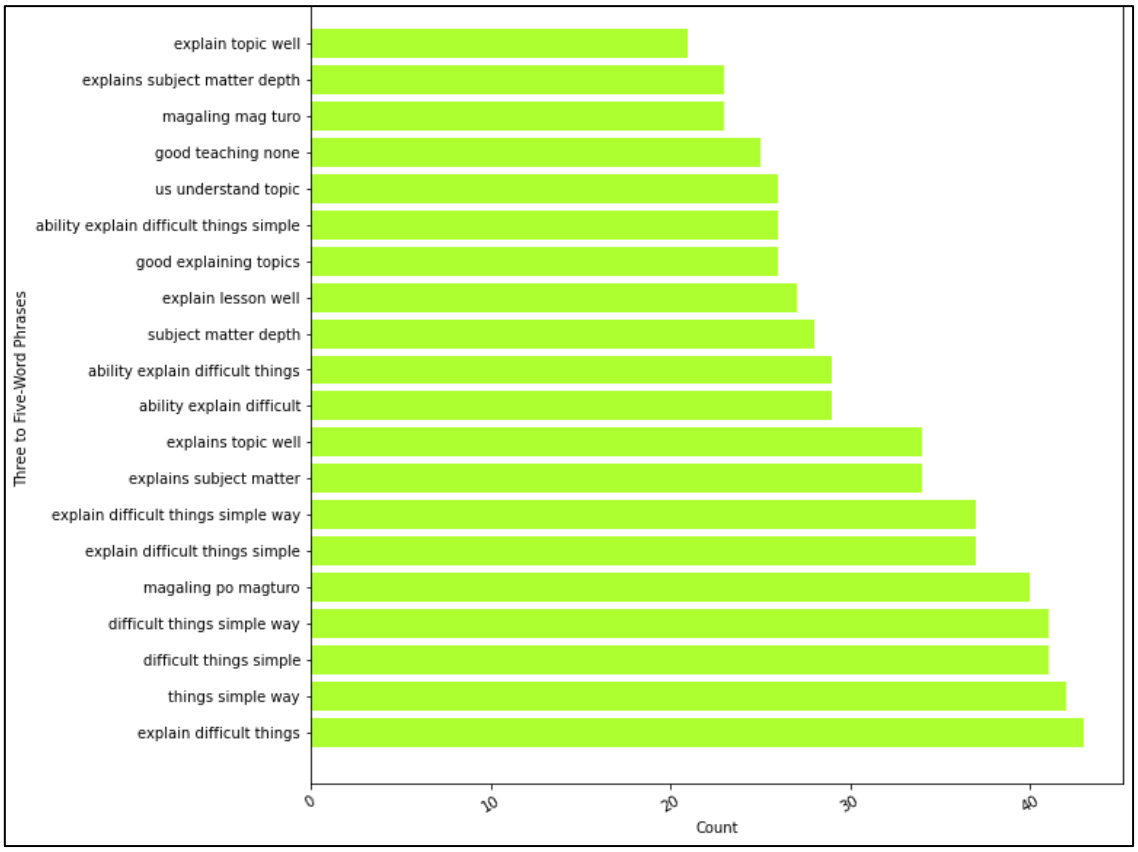

Fig. 6: Top 20 Most Frequent Phrases in Student Responses on Instructional Design/Delivery

Figure 6 shows most frequent phrases in comments describing instructional design/delivery are "*simply explain difficult things*", "*magaling magturo*", "*explains topic well*", "*explains subject matter with depth*", "*explain lesson well*" and "*explain the topic well*". Meanwhile, Figure 7 illustrates the most frequent phrases that were found in students' comments about assessment. Some of these are "*gives enough time*", "*mahirap magbigay quiz*", "*strict quizzes exams*", "*help us improve*" and "*lack feedback works*".

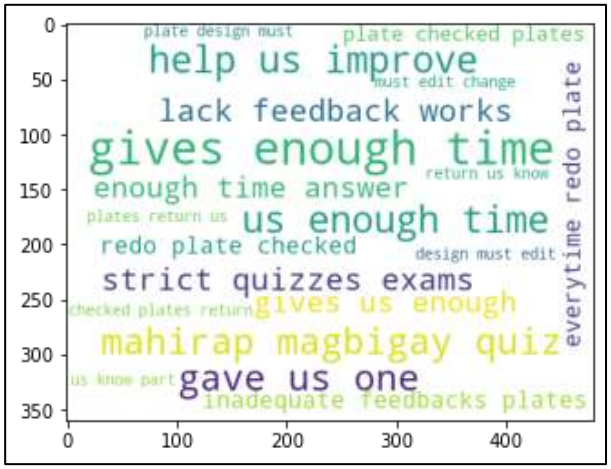

Fig. 7: Top 20 Most Frequent Trigrams from Comments on Student Assessment

Overall, the frequent phrases in students' responses indicate generally positive feedback on content and instructional design/delivery. Though there were few frequent phrases on negative comments describing personality/behavior, more positive frequent phrases were seen. Frequent phrases on feedback in student assessment are generally negative, though there were also few positive phrases. Meanwhile, the feedback on the use of technology is mostly describing the internet connection of teachers and students.

4 Conclusion

This study found that ensemble applied with Ngram vectorization is preferred over the base models in terms of sentiment classification while ensemble applied with Count vectorization is preferred in terms of topic classification. It can be concluded in this case that an ensemble of machine learning algorithms performs better than individual base models in sentiment analysis of teaching performance. Pie chart, bar graphs, and wordcloud were found to be comprehensible techniques to provide visualization of students' sentiments in online teaching performance.

In conclusion, this study holds substantial significance on multiple fronts. Firstly, it sheds light

on the nuanced sentiments of students towards faculty online teaching performance within a campus of a state university in the Philippines. Secondly, it distinguishes itself by employing a unique approach, utilizing bigrams and trigrams for sentiment visualization, thereby incorporating a more comprehensive analysis that integrates various aspects of teaching. Third, it presents a method in evaluating a sentiment classifier on the unbalanced dataset. Lastly, the study contributes valuable insights by presenting the outcomes of machine learning algorithms and vectorization techniques, providing a foundation for potential adaptations and enhancements in future research endeavors.

Though this study presented prominent machine learning algorithms used in sentiment analysis, the study is limited to the use of supervised and semi-supervised machine learning and basic vectorization techniques only. Neural networks for deep learning and more advanced techniques such as Word2Vec, Global Vectors and FastText may be considered in future studies.

References:

[1] Aristovnik, A., Karampelas, K., Umek, L. and Ravšelj, D. (2023). Impact of the COVID-19 pandemic on online learning in higher education: a bibliometric analysis. *Front. Educ.* 8:1225834. doi: 10.3389/feduc.2023.1225834

[2] Dolianiti, F.S., Iakovakis, D., Dias, S.B., Hadjileontiadou, S., Diniz, J.A., and Hadjileontiadis, L. (2019). Sentiment Analysis Techniques and Applications in Education: A Survey. *Technology and Innovation in Learning, Teaching and Education,* pp 412-427

[3] Abelito, J. T. and Baradillo, D.G. (2023). A Sentiment Analysis of College Students' Feedback on their Teacher's Teaching Performance During Online Classes. *International Journal of Multidisciplinary Educational Research and Innovation*, Volume 01 Issue 04

[4] Mamidted, A.D. and Maulana, S. S. (2023). The Teaching Performance of the Teachers in Online Classes: A Sentiment Analysis of the Students in a State University in the Philippines. *Randwick International of Education and Linguistics Science (RIELS) Journal*, Vol. 1, No. 1, March 2023, pp 86-95

[5] Rajput, Q., Haider, S. and Ghani, S., (2016). Lexicon-based sentiment analysis of teachers' evaluation. *Applied Computational Intelligence and Soft Computing*, 2016

[6] Madhuri, D. K. (2019). "A machine learning based framework for sentiment classification: Indian railways case study," *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075*, vol. 8, Issue-4, February 2019

[7] Lazrig, I. and Humpherys, S. (2022). Using Machine Learning Sentiment Analysis to Evaluate Learning Impact. *Information Systems Education Journal (ISEDJ)*. ISSN: 1545-679X

[8] Baragash, R.S., Aldowah, H. and Umar, I.N. (2022). Students' Perceptions of E-Learning in Malaysian Universities: Sentiment Analysis based Machine Learning Approach. *Journal of Information Technology Education*, Volume 21, 2022

[9] Lalata, J.A.P., Gerardo, B. and Medina, R. (2019). A sentiment analysis model for faculty comment evaluation using ensemble machine learning algorithms. In *Proceeding of 2019 International Conference on Big Data Engineering,* (pp. 68-73)

[10] Southern Regional Education Board (2006). *Checklist for Evaluating Online Courses*, [Online]. https://www.sreb.org/sites/main/files/file-attachments/06t06_checklist_for_evaluating-online-courses.pdf (Accessed Date: July 18, 2023).

[11] Llombart, O. R. *Using Machine Learning Techniques for Sentiment Analysis*, [Online]. https://ddd.uab.cat/pub/tfg/2017/tfg_70824/machine-learning-techniques.pdf (Accessed Date: July 19, 2023).

[12] Chakraborty, K., Bhattacharyya, S., Bag, R., and Hassanien, A.A. (2019). "Sentiment analysis on a set of movie reviews using deep learning techniques". *Social Network Analytics. Computational Research Methods and Techniques.* Pages 127-147. Available: https://doi.org/10.1016/B978-0-12-815458-8.00007-4.

[13] Wan, Y. (2015). *An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis*, [Online]. https://dalspace.library.dal.ca/bitstream/handle/10222/56328/Wan-Yun-MEC-ECMM-March-2015.pdf (Accessed Date: February 1, 2024).

[14] Data Science Wizards. (2023). *Guide to Simple Ensemble Learning Techniques*, [Online]. https://medium.com/@datasciencewizards/guide-to-simple-ensemble-learning-techniques-2ac4e2504912 (Accessed Date: February 1, 2024).

[15] Scikit-learn. *sklearn.metrics.cohen_kappa_score*, [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html (Accessed Date: July 10, 2023).

[16] Scikit-Learn. *sklearn.metrics.log_loss*, [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html (Accessed Date: July 20, 2023).

[17] Scikit-Learn. *sklearn.metrics.matthews_corrcoef*, [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html (Accessed Date: July 20, 2023).

[18] Sarlis, S. and Maglogiannis, I. (2020). On the Reusability of Sentiment Analysis Datasets in Applications with Dissimilar Contexts. *Artificial Intelligence Applications and Innovations*. 2020; 583: 409–418.

[19] Raza, G. M., Butt, Z.S., Latif, S. and Wahid, A. (2021). Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models. *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*.

[20] Kaplan, D. (2024). *Machine Learning 101: CountVectorizer Vs TFIDFVectorizer*, [Online]. https://enjoymachinelearning.com/blog/countvectorizer-vs-tfidfvectorizer (Accessed Date: February 4, 2024).

[21] Ogutu, R.V.A., Otieno C., and, Rimiru, R. (2022). Target Sentiment Analysis Ensemble for Product Review Classification. *Journal of Information Technology Research*, Vol. 15, Issue 1.

[22] Maitra, S. (2021). *Importance of Mathews Correlation Coefficient & Cohen's Kappa for Imbalanced Classes*, [Online]. https://sarit-maitra.medium.com/mathews-correlation-coefficient-for-imbalanced-classes-705d93184aed (Accessed Date: February 4, 2024).

[23] Niyaz, U. (2023). *Focal loss for handling the issue of class imbalance*, [Online]. https://medium.com/data-science-ecom-express/focal-loss-for-handling-the-issue-of-class-imbalance-be7addebd856 (Accessed Date: February 4, 2024).

Contribution of Individual Authors to the Creation of a Scientific Article

The sole author of this scientific article independently conducted and prepared the entire work from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This study was funded by the University Research Council of the Pangasinan State University.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.