Enhancing Sentiment Analysis of Textual Feedback in the Student-Faculty Evaluation using Machine Learning Techniques

***Caren A. Pacol [1], Thelma D. Palaoag [2]**
*[1]Department of Information Technology, Pangasinan State University, Urdaneta City, Philippines
[2]Department of Computer Science, University of the Cordilleras, Baguio City, 2600, Philippines

Abstract

Sentiment Analysis has been an interesting and popular research area encouraging researchers and practitioners to adopt this tool in various fields such as the government, health care and education. In education, instruction evaluation is one of the activities that sentiment analysis has served. Though, it is a common practice that educational institutions periodically evaluate their teachers' performance, students' comments which are rich in insights are not easily taken into account because of lack of automated text analytics methods. In this study, supervised machine learning algorithms were used. Experiments were conducted to evaluate base models employing naïve bayes, support vector machines and logistic regression in comparison to ensemble combining the three. Random forest, an ensemble learning algorithm was also experimented. Machine learning techniques such as term-frequency – inverse document frequency (TF-IDF) and ngram were also explored to devise a model with the highest possible accuracy. Results show that in this case, tf-idf vectorization does not show significant improvement in sentiment classification. On the other hand, ngram vectorization improve performance of base models and has potential to improve ensemble models. Random forest showed higher performance measures than base models and ensemble of the three base models. However, it did not outperform ngram combined with support vector machines.

In future work the model with highest accuracy found can be embedded in a sentiment analysis tool for students' feedback on teaching performance. More advanced transformation techniques and other ensemble techniques may be explored to further improve accuracy in sentiment classification.

Keywords: machine learning, sentiment analysis, teaching performance

1. Introduction

Sentiment Analysis has been an interesting and popular research area normally applied in reviewing products for business, understanding the mindset of people reading news and views

expressed by people in political debates. The ability of sentiment analysis to reveal opportunities to improve customer experiences, build better products and more has encouraged researchers and practitioners to adopt this tool in various fields such as the government, health care and education.

In education, instruction evaluation is one of the activities that sentiment analysis has served (Dolianiti et.al, 2019). It is important to monitor the satisfaction of students in the quality of instruction provided by their teachers. Though, it is a common practice that educational institutions periodically evaluate their teachers' performance, students' comments which are rich in insights are not easily taken into account because of lack of automated text analytics methods.

This study explored on employing machine learning and the different techniques to improve sentiment classification of students' feedback on teaching performance.

In the field of computer science, machine learning is the use of Artificial Intelligence (AI) in systems so that they become intelligent. The focus of machine learning is to produce algorithms that can possibly be used in AI applications in the real world. As enterprises are producing huge amount of data, it became indispensable to have machine learning techniques in place for discovering business intelligence from data for strategic decision making (Madhuri, 2019).

Machine learning belongs to supervised learning in general and text classification in particular. Thus it is also called as "Supervised Learning". Examples of techniques under supervised learning include Naïve Bayes, Support Vector Machine, Maximum Entropy, K-Nearest Neighborhood and Neural Networks (Chaudhari & Govilkar, 2015).

Machine learning allows capturing and examination of variations and meaning in client feedback across varied channels.

One great edge of machine learning is the capability to train the algorithms. Natural Language Processing (NLP) along with sentiment packages, sentiment corpora and other human-labelled sentiment rules are used to continuously improve algorithms thereby making them faster and more accurate (Whishworks, 2019).

Machine learning methods have limitations and, actually, can't work at a character level like humans. To improve the performance of the methods some transformations on the original data can be used that makes it easier to be processed by the machine learning methods. Each sentence is converted into a vector of features using count vector or term frequency-inverse document frequency (tf-idf) representation. Another transformation called n-grams makes the relation between words more explicit. Others are focused to generate new feature vectors that tries to represent the data in a more compact way and the rest are focused on modifying the original data to reduce the feature vector size (Llombart, 2017).

It is essential to analyze and interpret whether each students' textual feedback is positive or negative. When aggregated, it provides a picture of the overall satisfaction of student/s on teaching performance. However, doing this manually is a tedious task especially when a large

number of sentences are given. With this, machine learning can be applied to automatically classify positive, negative and neutral sentences.

At the very basic, machine learning algorithms can be categorized into supervised and unsupervised learning. In supervised learning, a target variable is to be predicted based on a given set of predictors. Inputs are mapped to expected outputs in a function using these group of variables. The model is trained until a target accuracy is attained using the training data. Meanwhile, no target variable is predicted in unsupervised learning. It is used for grouping population in what we call clusters and is widely used for subdividing clients in distinct groups for specific mediation (Ray, 2017).

In this study, supervised machine learning algorithms were used. Experiments were conducted to evaluate base models employing logistic regression, support vector machines and naïve bayes, in comparison to ensemble technique combining the three. Machine learning techniques such as term-frequency – inverse document frequency (TF-IDF) and n-gram were also explored. Use of random forest, a supervised ensemble learning algorithm was also experimented and compared to the other models.

Specific objectives of this study are: (a) evaluate the base models and ensemble model in classifying students' textual feedback (b) integrate and evaluate tf-idf and n-gram techniques to base models (c) evaluate model when tf-idf, n-gram and ensemble techniques are combined and (d) evaluate random forest ensemble when applied to sentiment classification.

2. Methodology

This chapter focuses on the discussion about the methods used in the process of the study.

2.1 Data Gathering

Teaching performance data were sourced from nine (9) campuses of Pangasinan State University (PSU). The data were taken from summary results of faculty teaching performance evaluation questionnaires of 367 faculty members. At present, PSU utilizes a Faculty Teaching Performance Evaluation composed of four (4) areas of evaluation namely commitment, knowledge of subject matter, teaching for independent learning and management of learning. There are key indicators in each area wherein students are allowed to rate their faculty from 1 to 5 (5 being the highest and 1 being the lowest). A section for narrative comments labelled Strengths and Weaknesses is in the latter portion of the evaluation form. The results were summarized and mean rating of each faculty member was calculated. Listing of comments on strengths and weaknesses are also summarized for every faculty.

2.2 Data Preparation

After data gathering, the comments were consolidated in an Excel file. The data were manually cleaned correcting misspelled words. A total of 9140 sentences were manually cleaned and

labelled with positive (1), negative (-1) and neutral (0) for the training dataset. Meanwhile, 1827 sentences were prepared and labelled for the test set.

2.3 Data Pre-processing

The data preparation is then followed by pre-processing wherein special characters were removed, multiple spaces were substituted with single space, prefixes were removed, all text were converted to lowercase and stop words were removed. The (Natural Language Toolkit) NLTK package in python was used in the pre-processing.

2.4 Training the Model

After pre-processing, the training data were used to train the model in classifying sentences. Three machine learning algorithms naïve bayes, support vector machines and logistic regression were used as base models.

2.5 Testing the Model and Measuring Accuracy

Then, the trained model is evaluated on the test set and accuracy is calculated using accuracy_score() function or classification_report() function of sklearn.metrics in python.

2.6 Applying TF-IDF, ngrams and Ensemble machine learning techniques in sentiment classification

Next is the conduct of experiment on improving the sentiment classifier by employing tf-idf, ngrams and ensemble techniques.

Ngrams and TF-IDF are approaches in text vectorization. Vectorization is the transformation of text into a meaningful vector or array of numbers a machine can understand. Count vector was used in the base models. In the count vector, the sentence is represented by the words and the number of occurrences of each word in the document generating a bag of word counts for each sentence (Llombart, 2017). TF-IDF and ngrams vector instead of count vector were separately used in two experiments.

In the base models, each sentiment is represented as binary vector. With TF-IDF, more information can be encoded into the vector. Term Frequency-Inverse Document Frequency (Tf-Idf) evaluates the significance of a word in a body of text data. Term frequency (Tf) measures the similarity among documents while Inverse document frequency (Idf) measures the importance of term. So, the multiplication of Tf and IDF of a word produces the frequency of this word in the document multiplied by uniqueness of the word (Chakraborty, 2019). TF is computed using Eq. 1. If term frequency is *Tf(wi, D)* and document frequency is *Df(wi)*. Then, from *Df(wi)*, inverse document frequency *Idf(wi)* is calculated using Eq. 2.

Tf(wi, D) = (# of times term w appears in document D) / (Total # of terms in document D) (1)

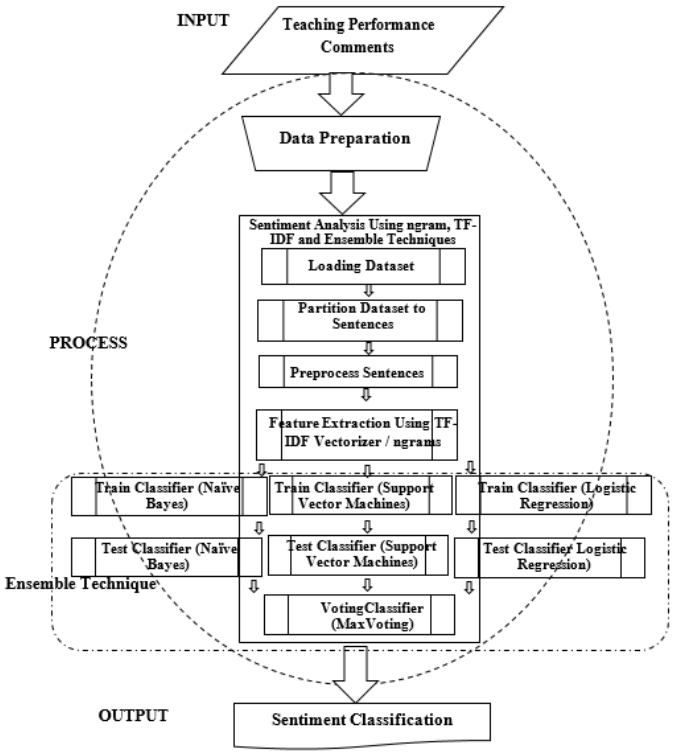
$$Idf(wi) = \log_e (\text{Total \# of documents} / Df(wi)) \quad (2)$$

The Tf-Idf of feature *wi* for document D is then calculated as the product: $Tf(wi,D) \cdot Idf(wi)$. The words with high Tf-Idf scored in a document are frequently occurred in that document and deliver the most important facts about the document (Chakraborty, 2019).

The ngrams are collections of words grouped by 1, in unigrams and 2, in case of bigrams and so on. For example, the sentence "Im not well" is converted to vector ("Im", "not", "well") in unigrams and ("Im not", "not well") in bigrams. In one experiment, ngram vector with ngram range set to 1 and 2 was used, which means sentences were converted to unigram and bigram vectors.

In another experiment, instead of using single models, they are combined in one model using Max Voting ensemble. Generally, the max voting is utilized for problems on classification. This approach utilizes several models in making predictions. Each model's prediction is called a 'vote'. The predictions which are from the majority of the models are used as the final prediction. Then, classification report and accuracy score of the new model were generated to compare its performance with the base models. Experiments were also conducted to evaluate combination of ngram and ensemble as well as tf-idf and ensemble. Fig. 1 shows the required input, processes to perform and the expected output in the ensemble model.

Figure 1: Sentiment classification using ngram, tf-idf and ensemble techniques



Then, ensemble learning algorithm called random forest was utilized. This ensemble learning algorithm produces decision trees using the data samples. Then it gets each of these decision trees' prediction. Finally, chooses the best solution through voting. Random forest algorithm is considered superior compared to single decision tree. This is because it avoids overfitting through getting the mean of results (Mishra, 2019).

3.7 Evaluating the Classification Model

Confusion matrix and classification report were produced to assess the classification model. The confusion matrix allow users to view a summary of prediction results identifying the number of correctly and incorrectly classified sentences in each class. To further analyze the performance of the sentiment classifier, the classification report shows the weighted average and macro-average results as well as precision, recall, f1-score and accuracy for each class. To get the weighted average, individual true positives, false positives and false negatives are totalled for the various classes and applied to get the statistics. Macro-average is calculated the average of the precision and recall of the system on different classes. Computations for precision, recall and F1 score in each class are in Eq. 3, Eq. 4 and Eq. 5 while overall accuracy is in Eq. 6.

Precision provides the percentage of correct positive predictions (Shung, 2018). It is calculated utilizing Eq. 3.

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives}) \quad (3)$$

Recall answers what percent of the positive cases were caught (Shung, 2018). It is calculated using Eq. 4.

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (4)$$

F1 score in Eq. 5 is used to seek balance between precision and recall (Shung, 2018).

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (5)$$

Accuracy measures how many are correct predictions among total predictions and is calculated with Eq. 6.

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) / (\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}) \quad (6)$$

3. Results and Discussions

This section present and discuss the results found in this study.

The test set consists of 1015 instances in class positive (1), 745 in class negative (-1) and 67 in class neutral (0). Macro and weighted averages of scores for precision, recall and F1 were calculated in the classification report. Since there is imbalance in the number of instances in each class, the weighted average results were considered and presented in Tab.1, Tab. 3 and Tab. 4.

Table 1: Performance of base and ensemble models

METRIC (Weighted average)	**BASE MODELS**			**ENSEMBLE MODELS**	
	LOGISTIC REGRESSION (LR)	**NAÏVE BAYES (NB)**	**SUPPORT VECTOR MACHINES (SVM)**	**ENSEMBLE (LR + NB + SVM)**	**RANDOM FOREST**
Accuracy	0.89	0.86	0.93	0.90	**0.96**
Precision	0.90	0.86	0.93	0.90	**0.96**
Recall	0.89	0.86	0.93	0.90	**0.96**
F1 Score	0.88	0.85	0.93	0.88	**0.96**

It can be observed in Tab. 1 that among the base models, Support Vector Machines gave the highest results in all of performance metrics. It is also shown ensemble technique (LR + NB +SVM) does not result to significant improvement in accuracy and other weighted average performance metrics while random forest yielded higher results of 0.96.

Predicting positive sentiments is necessary to see best practices of teachers. More so, predicting negative sentiments is crucial in this case since this require intervention measures to address the concern of students. Thus, higher value of recall in class -1 (pertaining to negative sentences) is desired. Recall in this context is the actual negative sentences that got predicted correctly in the case of class -1 and the actual positive sentences that got predicted correctly in the case of class 1. Tab. 2 indicates that random forest yielded higher recall in each of the three classes than the base models and the LR+NB+SVM ensemble model. However, recall for class -1 (negative sentences) is lower as compared to class 1 (positive sentences).

Results of experiments on tf-idf vectorization reveal that there is no significant improvement in the performance of base models. On the other hand, ngram vectorization gives significant improvement in all of the base models as shown in Tab. 3. Ngram + Support Vector Machines came out with the highest results of 0.98 in accuracy, precision, recall and F1 score.

Table 2: Recall of base and ensemble models

METRIC	**BASE MODELS**			**ENSEMBLE MODELS**	
	LOGISTIC REGRESSION (LR)	**NAÏVE BAYES (NB)**	**SUPPORT VECTOR MACHINES (SVM)**	**ENSEMBLE (LR + NB + SVM)**	**RANDOM FOREST**
Recall (Class -1)	0.88	0.85	0.91	0.88	**0.94**
Recall (Class 0)	0.19	0.13	0.58	0.09	**0.76**
Recall (Class 1)	0.95	0.93	0.97	0.96	**0.99**

Table 3: Performance of the ngram and base models

METRIC (Weighted average)	LOGISTIC REGRESSION	NGRAM + LOGISTIC REGRESSION	NAÏVE BAYES	NGRAM + NAÏVE BAYES	SUPPORT VECTOR MACHINES	NGRAM+ SUPPORT VECTOR MACHINES
Accuracy	0.89	0.97	0.86	0.90	0.93	**0.98**
Precision	0.90	0.97	0.86	0.91	0.93	**0.98**
Recall	0.89	0.97	0.86	0.90	0.93	**0.98**
F1 Score	0.88	0.97	0.85	0.89	0.93	**0.98**

Table 4: Performance of ensemble models with tf-idf and ngrams

METRIC (Weighted average)	ENSEMBLE (LR + NB + SVM)	TF-IDF + ENSEMBLE (LR + NB + SVM)	NGRAM + ENSEMBLE (LR + NB + SVM)	RANDOM FOREST	TF-IDF + RANDOM FOREST	NGRAM + RANDOM FOREST
Accuracy	0.90	0.89	0.91	**0.96**	0.95	**0.96**
Precision	0.90	0.89	0.92	**0.96**	0.95	**0.96**
Recall	0.90	0.89	0.91	**0.96**	0.95	**0.96**
F1 Score	0.88	0.87	0.90	**0.96**	0.95	**0.96**

Findings in the experiments on TF-IDF + ensemble model did not show improvement in performance of LR+NB+SVM and random forest. Meanwhile, ngram contributed improvement in the performance of LR+NB+SVM but not in random forest.

4. Conclusion and Future Works

This paper presents the results of experiments in machine learning algorithms and techniques applied in sentiment analysis of students' feedback on teaching performance. Results show that in this case, tf-idf vectorization does not show significant improvement in sentiment classification. On the other hand, ngram vectorization improve performance of base models and has potential to improve ensemble models. Ngram when combined with support vector machines yielded accuracy of 0.98 thereby increasing the accuracy and other performance metrics by 5%. Random forest showed higher performance measures than base models and ensemble of the three base models. However, it did not outperform ngram combined with support vector machines.

In future work the model with the highest performance measures can be used to devise a sentiment analysis tool for students' feedback on teaching performance. More advanced transformation techniques and other ensemble techniques may be explored to further improve accuracy in sentiment classification.

References

Chakraborty, K., Bhattacharyya S., Bag, R. and Hassanien, A.A. (2019). "7 - 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

Chaudhari, M. and Govilkar S. (2015). "A Survey of Machine Learning Techniques for Sentiment Classification," *International Journal on Computational Science & Applications (IJCSA)* vol.5, No.3, June 2015

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. International Conference on Technology and Innovation in Learning, Teaching and Education TECH-EDU 2018: Technology and Innovation in Learning, Teaching and Education pp 412-427

Llombart, O.R. (2017). Using machine learning techniques for sentiment analysis. [Online]. Available: https://www.semanticscholar.org/paper/Using-machine-learning-techniques-for-sentiment-Llombart/c6c19c0e9acea6372c93a7dcedb74883aa0b3580

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

Mishra, A.D. (2019). Ensemble learning and random forest. [Online]. Available: https://medium.com/datadriveninvestor/ensemble-learning-and-random-forest-7430ebf3da7e

Whishworks (January 2019). Machine learning in sentiment analysis. [Online]. Available: https://www.whishworks.com/blog/big-data/machine-learning-in-sentiment-analysis

Ray, S. (2017). Commonly used machine learning algorithms (with Python and R Codes). [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Shung, K. P. (2018). Accuracy, precision, recall or f1?. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9