Benchmarking Classical Machine Learning Models for Sentiment Classification in Students' Faculty Evaluations: A Comparative Study

Aaron Paul M. Dela Rosa *, Renato L. Adriano II, and Lilibeth G. Antonio

College of Information and Communications Technology, Bulacan State University, City of Malolos, Philippines
Email: aaronpaul.delarosa@bulsu.edu.ph (A.P.M.D.R.); renato.adriano@bulsu.edu.ph (R.L.A.); lilibeth.antonio@bulsu.edu.ph (L.G.A.)
*Corresponding author

Manuscript received October 1, 2025; revised October 30, 2025; accepted February 27, 2026; published July 17, 2026

***Abstract*—Sentiment analysis has emerged as a complementary aid to opinion extraction in textual information like academic faculty evaluations. Selecting the best machine learning algorithm involves balancing accuracy and computational complexity. This work compares classical machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Random Forest, and k-Nearest Neighbors (k-NN), for classifying student evaluations into negative, neutral, and positive sentiment. Training and testing were performed on labeled data, and performance was evaluated using precision, recall, F1-score, overall accuracy, a confusion matrix, and computational time. Random Forest was most accurate (99%) but took maximum computational time (4.575s). SVM came in a close second at 97%, with negligible computational cost, making it a formidable contender in applications where efficiency is a priority. Logistic Regression came in next at 94%, while Naïve Bayes was quickest (0.019s) but slightly less accurate (93%). k-NN performed worst at 88% and struggled to discriminate between overlapping sentiment classes. Confusion matrix analysis pointed out that Random Forest and SVM reduced misclassification primarily between neutral and positive sentiments. Analysis points out a trade-off between predictive accuracy and computational efficiency. Random Forest is superior for high-stakes sentiment classification, where predictive accuracy is prioritized above all else. Naïve Bayes and SVMs are preferred choices for applications that require real-time responses or where computational power is a limiting factor. Additional research should incorporate deep learning architectures and broader efficiency metrics, such as memory usage and energy consumption, to solidify applicable use cases for sentiment classification in education.**

***Keywords*—computational efficiency, faculty evaluations, machine learning, Random Forest, sentiment analysis, Support Vector Machine (SVM), text classification**

I. Introduction

Sentiment analysis, or opinion mining, is a central subfield of Natural Language Processing (NLP) that involves extracting and classifying subjective information from text [1, 2]. It has become a standard across a variety of applications, ranging from business intelligence and social media analysis to healthcare and education, and has been employed to produce actionable insights for decision-making [3, 4]. With the growing volume of unstructured data generated across online media, the demand for effective sentiment classification methods has increased exponentially, particularly in applications where institutional policy is informed by stakeholder feedback.

Sentiment analysis has emerged as a valuable tool in education to discern students' perceptions, satisfaction levels, and learning experiences [5, 6]. Student ratings of faculty work are especially desirable, as they provide direct measurement of teaching quality and course conduct, thereby facilitating institutional quality assurance and policy planning. However, manually interpreting such evaluations is a time-consuming, subjective process and, consequently, has led to automated sentiment analysis being sought after [7, 8].

Though much has been done in sentiment classification research, most work has been motivated by application use cases in commercial product reviews, movie databases, or large social media datasets [9, 10]. Considering other applications, relatively few works have addressed academic applications, and available works often employ deep learning-based strategies that require substantial computational power and large labeled datasets [11]. Although deep learning-based strategies are efficient, they may not always be available or interpretable in institutional settings where data are sparse, and openness is central to educational decision-making [12].

Machine learning has become a desirable method for sentiment classification since it can uncover patterns in labeled data. Traditional or "classical" machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes have long demonstrated outstanding performance in categorical classification problems since they can be interpreted readily, process data efficiently in a moderate amount of computer memory, and have moderate computational complexity [13, 14]. More recent ensemble-based algorithms, such as Random Forests, and instance-based algorithms, such as k-Nearest Neighbors (k-NN), have since demonstrated comparable performance across a wide range of sentiment analysis applications [15, 16]. These algorithms fundamentally differ in their learning strategies: Logistic Regression assumes linear decision boundaries, SVMs achieve class separation with hyperplanes, Naïve Bayes makes assumptions about conditional independence, Random Forests use ensemble decision trees to uncover nonlinear relationships, and k-NN uses distance-based similarity metrics to classify new instances. A comparative investigation of these algorithms is thus needed to discern their relative strengths in educational evaluation applications.

Although recent advances in NLP have been dominated by Transformer-based architectures, their adoption in educational analytics remains limited due to computational requirements, data volume constraints, and reduced interpretability. In institutional settings, such as

doi: 10.18178/ijiet.2026.16.7.2647

student-faculty evaluations, where datasets are typically moderate in size and transparency is essential for administrative decision-making, classical machine learning models remain practical and widely deployed solutions. Consequently, establishing robust empirical benchmarks for these models within authentic educational contexts remains a relevant and necessary research endeavor.

Even though sentiment analysis has gained wide acceptance in education, current work tends to focus on a few models or describe deep learning strategies without fully benchmarking classical algorithms against a range of algorithms [17, 18]. Though deep learning strategies such as recurrent and transformer-based networks have achieved mainstream popularity, classical algorithms remain highly applicable in realistic education analytics due to their low computational overhead, explainability, and suitability for smaller datasets [9, 19]. However, few comparative studies have fully evaluated trade-offs among classical algorithms in terms of accuracy, robustness, and computational efficiency using student-faculty evaluation data.

This work fills this gap by conducting a comparative assessment of the five most popular classical machine learning algorithms, including Logistic Regression, Support Vector Machines, Naïve Bayes, Random Forest, and k-Nearest Neighbors, for sentiment classification of student-faculty evaluation comments. By assessing the calibration and computational efficiency of these algorithms using standard performance metrics, this work identifies which algorithm performs best in sentiment classification in educational contexts. Its findings will not only be instrumental in method development in educational text mining but also become part of a broader evidence-based decision-making debate in higher education.

The general aim of this study is to benchmark and comparatively evaluate the classical machine learning model performance of Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Random Forest, and k-Nearest Neighbors, in sentiment classification on students' faculty evaluation data. Specific objectives include: (1) To evaluate the classification performance of models on standard metrics such as precision, recall, F1-score, and computation time, (2) To compare and assess the accuracy and capacity to handle imbalanced data of the five models using confusion matrices, and (3) To identify the best-suited model for sentiment classification in educational evaluation applications and propose applications in future decision-making in education.

II. Related Work

Sentiment analysis, or opinion mining, has emerged as a powerful analytical tool for analyzing textual data across a range of application contexts, including education. Sentiment analysis has numerous applications in higher education to transform unstructured feedback, such as faculty evaluations, student commentaries, and course reviews, into actionable feedback [20]. Textual comments provide richer information about the quality of teaching, the learning process, and student engagement than numeric rating scales [17]. It assists educational leaders in better identifying pedagogic concerns and using evidence-based quality improvement strategies.

Latest studies have emphasized the importance of sentiment analysis in educational contexts. For instance, Kastrati *et al*. [17] used machine learning-based sentiment analysis on students' course feedback and showed how it can be used to identify both levels of satisfaction and where improvement is required. Again, Abdi *et al*. [18] utilized deep learning algorithms to classify educational sentiments and emphasized the importance of sophisticated computational methodologies. Although sophisticated deep learning techniques have become increasingly common, classical machine learning algorithms remain important owing to their explainability, computational efficiency, and suitability for smaller educational datasets [19].

Classic machine learning algorithms have been extensively explored in sentiment analysis due to their simplicity, effectiveness, and comparable competence in text classification tasks. Among these algorithms, Logistic Regression (LR), Support Vector Machines (SVM), and Naïve Bayes (NB) are most widely used. Logistic Regression, although linear, has been observed to perform remarkably well on sparse, high-dimensional data such as text [14]. SVMs, due to their ability to produce optimal hyperplanes, have been highly successful in sentiment polarity detection tasks [13]. Naïve Bayes is valued for its probabilistic approach and computational efficiency, which enable it to perform effectively at a large scale in text mining [21].

Although these models have been studied extensively, they can work very differently on domain-specific datasets. For education, feedback is often contextual but sparse in numbers; hence, a systematic assessment across these models is needed to pinpoint those suited for use [17].

In addition to the classical trio SVM, LR, and NB, Random Forest (RF) and k-Nearest Neighbors (k-NN) remain two distinct yet operational paradigms within classical machine learning. Random Forest, a decision tree-based ensemble method that aggregates decision trees, can identify nonlinear associations and prevent overfitting through bootstrap aggregation [16]. Its flexibility across varying feature distributions and noise robustness make it a top contender for tasks such as sentiment analysis and text classification [15].

However, k-NN is a non-parametric, instance-based learning algorithm that makes classification predictions for new samples based on how closely they resemble labeled instances. Although computationally costly during inference, k-NN has been reported to achieve strong performance in sentiment analysis when combined with appropriate feature extraction techniques [19]. Reliance on similarity metrics further ensures a transparent interpretation of classification decisions, which is highly preferred in learning applications that prioritize interpretability.

Recent comparisons have shown that considering RF and k-NN alongside LR, SVM, and NB yields a more informative evaluation of conventional machine learning algorithms for text classification [16, 22]. Therefore, considering these in this work closes a gap in the literature by systematically comparing a broader range of conventional models in educational sentiment analysis.

Classification scheme evaluation requires rigorous testing using common metrics that capture both precision (the accuracy of positive instances) and recall (the completeness of positive identification). Sentiment analysis research often

reports precision, recall, and F1-score as key metrics [3, 10].

For example, Zhang *et al*. [4] highlighted that the F1-score provides a balanced assessment of classifier performance, especially when accuracy is a poor proxy due to imbalanced datasets. Tao *et al*. [8] went further and employed precision–recall analysis in educational feedback to reveal how model selection impacts both sensitivity to subtle sentiments and overall classification soundness.

Thus, using these commonly used metrics ensures a basis for comparison with previous work and enhances the credibility of the evaluation findings.

Aside from quantitative metrics, confusion matrices are central to model strength analysis, particularly identifying misclassification within sentiment labels (Positive, Negative, and Neutral). The literature shows that confusion matrices help identify model weaknesses, such as Naïve Bayes's tendency to overpredict the majority class or an occasional misclassification of neutral samples as positive in SVM-based systems [6, 11].

These diagnostic tools provide a deeper understanding of a classifier's relative strengths and enable clearer benchmarking. Whereas mislabeling sentiments could lead to misunderstandings about student satisfaction in an educational application use case, confusion matrix analysis ensures a rigorous methodology and accountability.

III. Materials and Methods

The methodology of this study was designed to systematically implement, evaluate, and compare five classical machine learning models, namely, Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Random Forest, and k-Nearest Neighbors, for sentiment classification on students' faculty evaluation data. The process encompassed dataset preparation, text preprocessing, model training, evaluation, and a structured comparative analysis.

A. Dataset Preparation

The dataset used in this study consists of textual feedback collected from official student-faculty course evaluation forms administered at a public higher education institution in the Philippines. The feedback was gathered across multiple undergraduate programs offered by the College of Information and Communications Technology from Academic Year (AY) 2021–2022 to AY 2023–2024. All student comments were written in English, the primary medium of instruction for institutional evaluations. Each AY, there were thousands of raw data points to process. Each record comprised a written comment provided by a student and its corresponding sentiment label, categorized as Positive, Negative, or Neutral. The raw dataset was extracted from institutional evaluation records and exported to a structured spreadsheet, with each row representing a unique student comment. To ensure balanced representation, exploratory data analysis was performed to examine class distribution. The number of comments for each sentiment polarity was balanced to avoid bias; e.g., more positive comments could bias the Positive sentiment polarity during model training. Each polarity has 1,042 comments to ensure dataset balance. In total, the dataset contains 3,126 data points. Since sentiment categories are often imbalanced in real-world educational contexts, stratified sampling was applied during the training-test split, ensuring that both the training and test sets maintained proportional representation of each sentiment category.

The dataset size reflects real-world institutional constraints, as student-faculty evaluations are typically collected on a per-term or per-program basis, limiting the feasibility of large-scale data aggregation.

The dataset was anonymized prior to analysis to protect student privacy, and no personally identifiable information was included. Ethical approval and institutional permission were obtained before data extraction, ensuring compliance with institutional data governance policies.

B. Data Preprocessing

Textual feedback often contains noise in the form of varied capitalization, special characters, filler words, and linguistic variations. To address these issues, a series of preprocessing steps was conducted. First, all text was converted to lowercase to ensure consistency in token matching. Punctuation marks, special characters, and numeric values were removed to reduce noise. The comments were then tokenized and split into words for further analysis. Stop words such as "the", "is", and "and" that bear no weight in sentiment identification were eliminated. Stemming and lemmatization were applied to normalize words to their roots, thereby reducing linguistic differences such as "teaching" and "taught" to "teach". Then the cleaned-up text was transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm, which both retains word frequencies and relative importance for discriminating sentiment across comments.

TF–IDF was selected as the feature representation method due to its interpretability, computational efficiency, and suitability for classical machine learning models. While distributed word embeddings such as Word2Vec and GloVe, as well as contextual embeddings derived from transformer-based models (e.g., BERT), provide richer semantic representations, they introduce higher computational complexity and reduced transparency. Given the study's objective of benchmarking classical models in resource-constrained educational settings, TF–IDF was deemed an appropriate and justifiable choice.

C. Model Training and Implementation

Selecting appropriate machine learning algorithms is required to achieve sound, interpretable results in applications involving sentiment classification. Five widely known and popular algorithms were selected in this work: Logistic Regression, Support Vector Machines (SVM), Naïve Bayes, Random Forest, and k-Nearest Neighbors (k-NN). These algorithms represent different methodological positions in text classification and thus provide a good basis for a comparative study. Table 1 presents why such algorithms were selected to be compared in this work.

Logistic Regression (LR) was chosen for its strong theoretical foundation in statistical learning and its strong performance on linearly separable text classification problems. Its comparable performance on high-dimensional text data has been upheld in recent studies, especially when model coefficient interpretation is deemed important [23, 24].

Table 1. Rationale for selecting machine learning models for sentiment classification

Model	Rationale for Selection	Supporting References
Logistic Regression (LR)	A commonly used linear model due to its simplicity, interpretability, and satisfactory performance on sparse high-dimensional data such as TF-IDF features. LR tends to be a good baseline for sentiment analysis tasks.	Wang *et al*. [23]; Das *et al*. [24]
Support Vector Machine (SVM)	Most notable for its effective operation in high-dimensional feature spaces, SVM optimally constructs hyperplanes to classify. SVMs do not overfit to sparse data like text and are among the most consistent performers in sentiment classification.	Ahmad *et al*. [25]; Mao *et al*. [26]
Naïve Bayes (NB)	A probabilistic classifier that assumes conditional independence among features, rendering it computationally efficient and successful in handling text data. NB usually performs pretty well on small to medium collections and is a good baseline classifier in particular for sentiment analysis.	Kowsari *et al*. [27]; Ashbaugh & Zhang [28]
Random Forest (RF)	An ensemble learning method where multiple decision trees are combined to generalize better and reduce overfitting. RF can model nonlinear relationships and generate feature importance, making it suitable for text classification use cases that require robustness and interpretability.	Bahrawi [29]; Elfaik & Nfaoui [30]
k-Nearest Neighbors (k-NN)	An instance-based, non-parametric process making classification decisions based on similarity to training instances. Although computationally costly on large datasets, k-NN is a desirable alternative to parametric models and is interpretable in classification problems.	Wang *et al*. [31]; AlGhamdi & Khatoon [32]

Support Vector Machines (SVMs) remain a gold-standard algorithm in text mining, particularly for sentiment analysis, because they can operate effectively on sparse, high-dimensional data typical of textual inputs. The literature has extensively validated that SVMs are highly effective at optimizing the separation between classes using hyperplanes [25, 26].

Naïve Bayes (NB), a Bayes' theorem-driven probabilistic classifier, was given consideration owing to its simplicity, efficiency in computation, and satisfactory baseline performance in sentiment classification. Despite having a conditional feature independence assumption, NB still performs well in sentiment analysis, particularly for short messages and highly imbalanced data [27, 28].

Random Forest (RF) was incorporated to achieve an ensemble-based solution that leverages the collective strength of many decision trees to yield high accuracy and robustness against overfitting. Recent research has shown that Random Forests are extraordinarily successful in text-based sentiment classification, especially for large, noisy data sets, while retaining predictive power and model interpretability [29, 30].

Finally, k-Nearest Neighbors (k-NN) was selected to showcase instance-based learning, making predictions for new data points based on how close they are to labeled samples in the feature space. Although computationally intensive for large datasets, k-NN has been shown to accurately classify sentiment when combined with optimized distance metrics and dimension reduction [31, 32].

These five models provide a reasonable coverage of statistical, probabilistic, ensemble, and instance-based paradigms of learning. Such diversity enables a comprehensive benchmark study that ensures the evaluation covers predictive performance and interpretability across a variety of modeling strategies.

All experiments were conducted using Python (Scikit-learn library) on Google Colab's standard CPU runtime. The local client device was a Windows 11 machine equipped with an 11th-generation Intel Core i5 processor and 16 GB RAM. No GPU acceleration was used. Execution times reported in this study represent end-to-end training and testing durations under this environment. The hyperparameters have been calibrated via grid search and cross-validation to achieve optimal performance. Two scenarios were created using a train-test split for better performance comparison. The train-test splits used to compare the model's performance were 80-20 and 70-30. Stratified sampling has been used to preserve class distribution within both subsets, thereby making the results more trustworthy and more likely to generalize. The distribution of each sentiment polarity around training and testing subsets is shown in Table 2.

Table 2. Dataset distribution for training and testing subsets

Sentiment Polarity	80–20 Split		70–30 Split	
	Train	Test	Train	Test
Negative	834	208	730	312
Neutral	833	209	729	313
Positive	833	209	729	313
Total	**2500**	**626**	**2188**	**938**

To ensure that the trained models do not exhibit bias toward a single sentiment polarity, an equal distribution was achieved through stratified sampling. On an 80–20 split, 2500 data points were used as the training subset, while the remaining 626 data points were used as the testing subset. On the other hand, a 70–30 split yields 2188 data points in the training set and 938 in the test set. With two compared scenarios, this will provide more insight into how the models will perform across different dataset splits.

D. Evaluation Metrics

The performance of the sentiment classification models in this study is evaluated using four core metrics: accuracy, precision, recall, and the F1-score. These are derived from the confusion matrix, which outlines correctly and incorrectly classified instances across sentiment categories [33]. These evaluation measures are widely recommended in NLP studies because they account for class imbalance, which is often prevalent in real-world sentiment data [34, 35].

The model's accuracy is the proportion of instances correctly classified relative to the overall dataset. Accuracy provides an overall measure of correctness but can be a misleading index in cases of class imbalance, where a large number of instances correspond to a specified class [36]. Precision determines how accurately a model classifies correctly and is relevant when false positives are costly. Recall is primarily about how accurately a model identifies all instances that correspond to true positives, ensuring that valuable categories, such as negative feedback, are not left behind [37]. Precision and recall provide complementary

views of model performance, with precision emphasizing correct positive predictions and recall emphasizing complete positive detection.

F1-score is calculated as a harmonic mean between recall and precision. It is desirable in natural language applications whose distributions may be imbalanced [30]. Precision and recall both may suggest trade-offs between two classes in imbalanced data sets. Thus, the F1-score provides a harmonic balance, making it a good measure across all classes.

With these classical metrics excluded, computational time was considered an additional measure of model performance. Computational time is the time it takes a given algorithm to complete both the training and testing phases. The reported execution time (in seconds) represents the combined duration of model training (fit) and prediction (predict) on the test set using precomputed TF-IDF features. It does not include preprocessing steps such as text cleaning, TF-IDF vectorization, train-test splitting, cross-validation, or hyperparameter tuning. All models were evaluated under identical computational conditions to ensure fair and consistent comparison across classifiers. It's a valuable measurement because extremely accurate predictive models will be of little help in implementing large-scale or time-sensitive applications if substantial resources or a long time are required to process them. Latest research emphasizes that if computational efficiency is considered alongside metrics based on accuracy, better overall benchmarking between models occurs, particularly between lighter algorithms such as Naïve Bayes and heavier algorithms such as Random Forests or Support Vector Machines [30, 38].

To assess the statistical robustness of model performance, five-fold stratified cross-validation was conducted on the training data. Mean accuracy and F1-scores, along with their corresponding standard deviations, were computed for each classifier. Additionally, paired statistical significance testing was performed between top-performing models using a paired t-test to determine whether observed performance differences were statistically significant.

By combining these evaluation metrics, the study ensures a rigorous, comprehensive assessment of sentiment classification performance, balancing predictive quality, robustness, and computational efficiency.

IV. Results and Discussion

A. Model Classification Performance Evaluation

To properly assess the performance of the five compared models, Precision, Recall, F1-score, Accuracy, and computational time were considered. They are widely regarded as reliable indicators in classification task evaluation, especially in imbalanced datasets, where a single measure such as accuracy can be misleading [39, 40]. Table 3 presents the classification performance metrics for the five compared models.

Table 3. Classification performance metrics of the five compared models

Train-Test Split	Model Performance Metrics		LR	SVM	NB	RF	k-NN
80–20 Split	Accuracy		0.94	0.97	0.93	0.99	0.88
	Precision	Negative	0.95	0.96	0.95	0.99	0.89
		Neutral	0.92	0.95	0.93	0.98	0.87
		Positive	0.96	1.00	0.91	1.00	0.88
	Recall	Negative	0.96	1.00	0.94	1.00	0.85
		Neutral	1.00	1.00	0.9	1.00	0.97
		Positive	0.87	0.91	0.93	0.97	0.82
	F1-score	Negative	0.96	0.98	0.94	0.99	0.87
		Neutral	0.96	0.97	0.92	0.99	0.92
		Positive	0.91	0.95	0.92	0.99	0.85
	Time (s)		0.78	0.09	0.02	4.58	0.07
70–30 Split	Accuracy		0.93	0.97	0.92	0.99	0.84
	Precision	Negative	0.94	0.96	0.94	0.99	0.85
		Neutral	0.90	0.95	0.94	0.97	0.84
		Positive	0.95	1.00	0.89	1.00	0.83
	Recall	Negative	0.95	0.98	0.94	0.98	0.79
		Neutral	0.96	1.00	0.89	1.00	0.95
		Positive	0.88	0.92	0.94	0.98	0.79
	F1-score	Negative	0.94	0.97	0.94	0.99	0.82
		Neutral	0.93	0.97	0.91	0.98	0.89
		Positive	0.91	0.96	0.91	0.99	0.81
	Time (s)		3.08	0.24	0.18	8.80	0.15

Under the 80–20 split, Random Forest achieved the highest accuracy (0.99), followed by SVM (0.97), LR (0.94), NB (0.93), and k-NN (0.88). A similar ranking was observed for the 70–30 split, with Random Forest maintaining an accuracy of 0.99, SVM at 0.97, LR at 0.93, NB at 0.92, and k-NN at 0.84.

For the 80–20 split, Random Forest recorded precision values of 0.99 for the Negative class and 1.00 for both Neutral and Positive classes. SVM achieved precision values ranging from 0.95 to 1.00 across all classes. Logistic Regression and Naïve Bayes produced precision values above 0.90 for all sentiment categories. k-NN yielded lower precision scores, particularly for the Neutral (0.87) and Positive (0.88) classes.

Under the 70–30 split, Random Forest maintained precision values between 0.97 and 1.00 across all classes. SVM again demonstrated high precision, with values of 0.96, 0.95, and 1.00 for Negative, Neutral, and Positive classes, respectively. Precision values for k-NN declined across all classes, with the lowest observed in the Positive class (0.83).

Recall results under the 80–20 split indicate that Random Forest achieved recall values of 1.00 for the Negative and Neutral classes and 0.97 for the Positive class. SVM recorded recall values of 1.00 for both Negative and Neutral classes and 0.91 for the Positive class. Logistic Regression and Naïve Bayes showed recall values ranging from 0.87 to 1.00 across

classes. k-NN obtained recall values of 0.85, 0.97, and 0.82 for the Negative, Neutral, and Positive classes, respectively.

For the 70–30 split, Random Forest achieved recall values of 0.98-1.00 across all classes. SVM produced recall values of 0.98, 1.00, and 0.92 for Negative, Neutral, and Positive classes, respectively. k-NN exhibited the lowest recall values, particularly for the Negative and Positive classes (0.79).

Under the 80–20 split, Random Forest achieved F1-scores of 0.99 across all sentiment classes. SVM recorded F1-scores between 0.95 and 0.98, while Logistic Regression and Naïve Bayes obtained F1-scores ranging from 0.91 to 0.96. k-NN produced F1-scores between 0.85 and 0.92.

A similar pattern was observed under the 70–30 split, where Random Forest achieved F1-scores of 0.98–0.99 across all classes. SVM maintained F1-scores above 0.96 for the Negative and Neutral classes and 0.96 for the Positive class. Logistic Regression and Naïve Bayes exhibited F1-scores between 0.91 and 0.94, while k-NN recorded the lowest F1-scores, particularly for the Positive class (0.81).

Execution time measurements show that Naïve Bayes required the shortest processing time across both splits (0.02–0.18 s), followed by k-NN and SVM. Logistic Regression required moderate processing time (0.78–3.08 s). Random Forest exhibited the longest execution time, increasing from 4.58 s under the 80-20 split to 8.80 s under the 70-30 split.

The comparative evaluation of machine learning models reveals distinct performance patterns that are consistent with their underlying learning mechanisms. Ensemble-based and margin-based models demonstrated superior, more stable predictive performance compared to probabilistic and instance-based approaches, particularly in multi-class sentiment classification. These findings underscore the importance of model complexity and decision boundary optimization when working with sentiment-oriented textual data.

The consistently strong performance of the Random Forest classifier indicates its effectiveness in capturing non-linear feature interactions and mitigating overfitting through ensemble aggregation [29, 30]. Its robustness across different class labels suggests greater generalization across sentiment categories, including those with overlapping linguistic characteristics. Similarly, the strong performance of the Support Vector Machine reflects its ability to construct optimal separating hyperplanes in high-dimensional feature spaces, a property well-suited for text-based representations [25, 26].

In contrast, the comparatively moderate performance of Logistic Regression and Naïve Bayes suggests limitations in modeling complex class boundaries. While these models benefit from simplicity and computational efficiency, their assumptions of linear separability and conditional independence may limit their effectiveness in sentiment classification tasks where contextual dependencies and semantic overlap are prevalent [23, 28]. The consistently lower performance of k-Nearest Neighbors further highlights the sensitivity of distance-based methods to data distribution and sample density, particularly when the training data are reduced [31].

The computational analysis reveals a clear trade-off between predictive performance and execution cost. While ensemble methods incurred higher computational overhead, their performance gains suggest that such costs may be justified in applications where classification accuracy and reliability are prioritized. Conversely, Naïve Bayes remains suitable for scenarios requiring rapid model deployment or real-time inference, despite its comparatively lower predictive capability.

The comparative analysis of the 80–20 and 70–30 train-test splits indicates that model performance rankings remained consistent across both configurations, suggesting overall stability of the classifiers. However, the 80–20 split yielded marginally stronger and more stable performance across models, particularly for complex classifiers that benefit from larger training sets. This observation aligns with established machine learning practice, where increased training data generally enhances the learning of feature representations and class boundaries [39, 40].

Given these considerations, the 80–20 train-test split was selected for all succeeding analyses in this section. This split offers a balanced trade-off between training sufficiency and evaluation reliability, ensuring that models are trained on a larger portion of the dataset while retaining an adequate test set for unbiased performance assessment. Adopting a single, consistent split also enhances the clarity and comparability of subsequent results, particularly in extended analyses involving model comparison, error analysis, and potential optimization.

B. Model Robustness Using Confusion Matrices

Beyond the performance metrics used, confusion matrices offer valuable insights into how models classify and predict sentiment. Confusion matrices are used to present misclassification rates for models trained on the training subset and tested on the test subset. Figs. 1–5 present the confusion matrices for each model after training and testing on the test data subset.

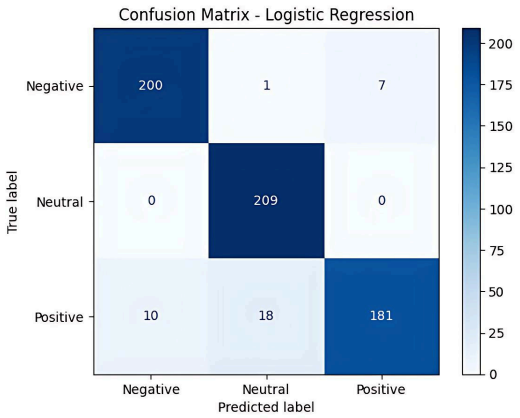

Fig. 1. Confusion matrix of logistic regression.

As shown in Fig. 1, the confusion matrix for Logistic Regression demonstrates high precision and recall across all three sentiment classes: 200 out of 208 negative reviews, 209 out of 209 neutral reviews, and 181 out of 209 positive reviews correctly classified. However, there are some weaknesses, particularly in distinguishing positive sentiments: 18 positive samples were misclassified as neutral, and 10 as negative. This indicates that while the model effectively captures neutral sentiment, it struggles to

capture the subtle lexical and semantic features that distinguish positive reviews. This challenge has been observed in prior work, where Logistic Regression was shown to perform consistently but sometimes underestimates positive polarity in nuanced sentiment tasks [41]. Despite these limitations, its overall balance across categories supports its use as a baseline model.

As shown in Fig. 2, the SVM matrix achieves nearly flawless classification, particularly for negative and neutral categories, misclassifying only one negative sample and correctly identifying all neutral samples. For positive reviews, 190 of 209 were correctly classified, though 9 were misclassified as negative and 10 as neutral. These results emphasize the strength of SVM in handling high-dimensional text data, where its ability to construct optimal hyperplanes maximizes class separation [42]. Its robustness is particularly evident in the minimal misclassification of neutral cases, a common challenge in sentiment analysis since neutral content often shares features with both positive and negative sentiments. Compared to Logistic Regression, SVM demonstrates stronger discriminative power, confirming its standing as one of the most reliable classifiers for sentiment tasks.

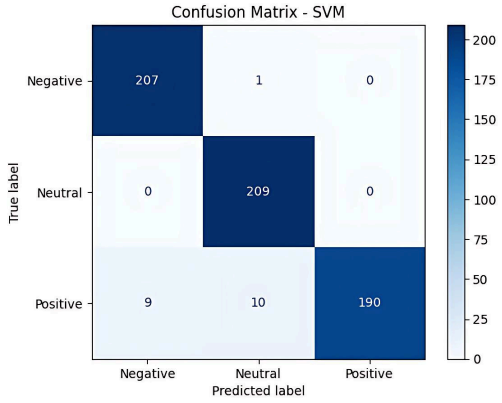

Fig. 2. Confusion matrix of support vector machine.

The Naïve Bayes confusion matrix in Fig. 3 highlights the model's relative strengths and limitations. While it correctly classified the majority of negative (196), neutral (189), and positive (195) samples, it showed noticeable confusion between the neutral and positive classes. For example, 12 neutral cases were misclassified as positive, and 11 positive cases were classified as neutral. This reflects the core limitation of Naïve Bayes: its assumption of conditional independence between features. Sentiment-laden words often co-occur and interact contextually, violating this assumption and resulting in weaker performance in distinguishing fine-grained emotional tones. This behavior has been confirmed in comparative studies, where Naïve Bayes often provides efficient but less context-sensitive predictions [43]. Nevertheless, its relatively low computational cost makes it useful for quick baseline modeling in sentiment tasks.

Fig. 4's Random Forest model confusion matrix has the highest performance, showing nearly perfect classification across all sentiment categories. Six out of a total of 626 samples were incorrectly classified, of which three positive samples were classified as neutral and another three as negative. Such impeccable performance demonstrates Random Forest's prowess as an ensemble method that aggregates many decision trees to reduce variance and overfitting while uncovering in-depth nonlinear relationships in text data. Its ability to handle both imbalanced and high-dimensional data makes it particularly suited for sentiment classification, as supported by recent ensemble learning studies [44]. Most notable is Random Forest's strength in achieving balanced overall accuracy across all sentiment classes, rather than biasing discrimination toward positive and neutral categories like Naïve Bayes or Logistic Regression.

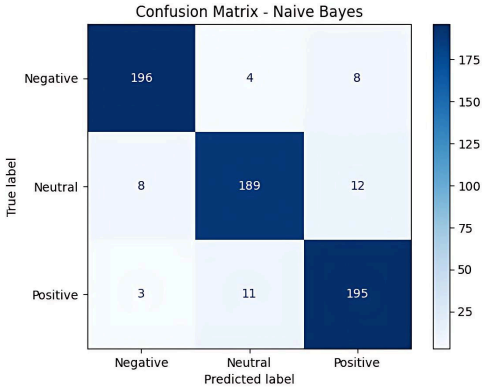

Fig. 3. Confusion matrix of Naïve Bayes.

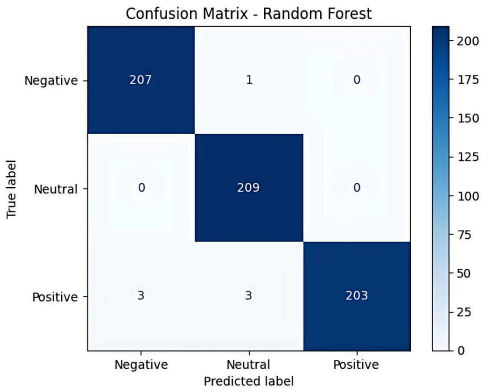

Fig. 4. Confusion matrix of random forest.

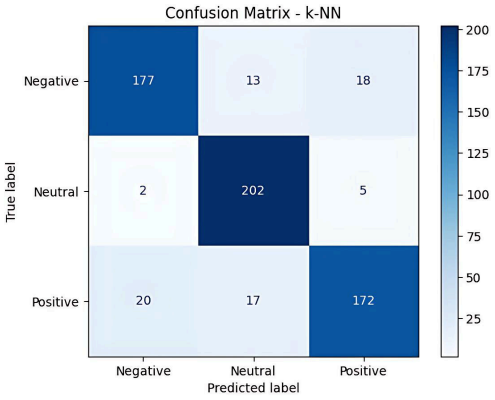

Fig. 5. Confusion matrix of k-Nearest neighbors.

Fig. 5's k-NN confusion matrix shows lower robustness than that of other models. Out of a total number of 208 negative cases, only 177 were predicted correctly, but 31 were predicted incorrectly, either as neutral or positive. Once again, 20 positive samples were incorrectly predicted as negative, and 17 as neutral, resulting in a total of 57 errors in the entire dataset. This behavior reflects k-NN's sensitivity to overlapping feature spaces, especially in high-dimensional

text vectors where class boundaries may not be distinct. Additionally, k-NN's reliance on distance metrics makes it prone to misclassifications when sentiment-bearing words are distributed across multiple classes. Previous studies have observed similar patterns, noting that while k-NN can effectively capture local patterns, it scales poorly in text sentiment tasks and is less robust under data imbalance [45]. Despite its limitations, k-NN remains valuable for small datasets or where interpretability of "neighbor-based" predictions is desired.

Across the five models, clear distinctions emerge in terms of classification robustness. Random Forest and SVM outperformed the others, achieving near-perfect accuracy and minimal misclassifications across categories. Logistic Regression also performed strongly, though with slight difficulty in differentiating positive sentiment. Naïve Bayes achieved competitive performance but struggled with distinguishing between neutral and positive cases, while k-NN showed the least robustness, with relatively higher misclassification rates. These findings align with prior research, which consistently shows that ensemble-based and margin-based classifiers (Random Forest, SVM) are superior for text sentiment classification compared to probabilistic (Naïve Bayes) and distance-based (k-NN) approaches [46, 47].

C. Model Selection for Sentiment Classification in an Educational Context

The models were compared not only based on accuracy and standard evaluation metrics, but also on computational efficiency. Fig. 6 illustrates the relationship between classification accuracy and training time for the five models: Logistic Regression, SVM, Naïve Bayes, Random Forest, and k-NN. This analysis of accuracy and computational time provides a holistic perspective on the suitability of the model for practical sentiment analysis applications.

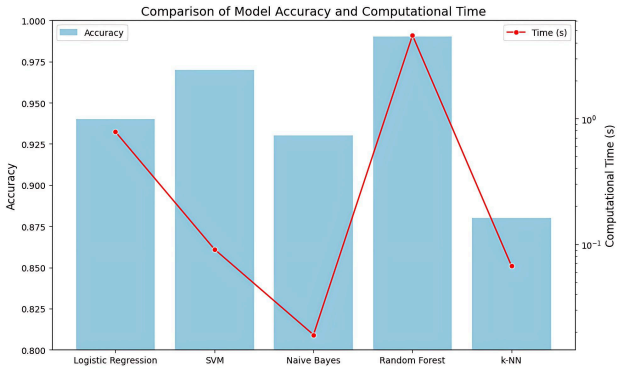

Fig. 6. Relationship between classification accuracy and computational time for the five evaluated machine learning models.

Fig. 6 shows that although Random Forest achieved the highest accuracy, it required the longest training time due to its ensemble structure of multiple decision trees. This trade-off aligns with the existing literature, which praises ensemble models for their predictive power but criticizes them for their computational overhead [44]. In contrast, Naïve Bayes recorded the fastest execution time while still maintaining reasonably high accuracy. Its efficiency is attributed to its simple probabilistic framework, which avoids iterative optimization and hyperparameter tuning, making it appealing for real-time applications despite its slightly weaker robustness [43].

SVM achieved competitive accuracy with moderate training time, reinforcing its balance between performance and efficiency. This result supports Khanam (2023), who emphasized SVM's effectiveness in sentiment analysis through optimal hyperplane separation. Similarly, Logistic Regression delivered reliable performance across categories, with shorter training times than ensemble methods, reaffirming its role as a strong baseline classifier for text-based tasks [41].

In contrast, k-NN was both computationally expensive and less accurate. This result is expected, as k-NN computes distances against the entire dataset during prediction, leading to high memory and time costs, especially in high-dimensional text representations [45]. Although its neighbor-based predictions are intuitive, this inefficiency limits its suitability for large-scale sentiment classification.

To ensure robust and reliable experimental findings, a stratified 5-fold cross-validation was performed for all classification models. Performance was evaluated using accuracy and macro-F1 score, and results were reported as mean ± standard deviation across folds. The inclusion of standard deviation enables assessment of performance stability and model generalization consistency.

Table 4. Cross-validation performance results

Model	Accuracy (Mean ± SD)	Macro-F1 (Mean ± SD)
LR	0.9568 ± 0.0082	0.9565 ± 0.0084
SVM	0.9750 ± 0.0090	0.9749 ± 0.0091
NB	0.9367 ± 0.0141	0.9369 ± 0.0138
RF	0.9933 ± 0.0035	0.9933 ± 0.0035
k-NN	0.9056 ± 0.0037	0.9049 ± 0.0034

As shown in Table 4, Random Forest achieved the highest performance with a Macro-F1 score of 0.9933 ± 0.0035, followed by Support Vector Machine (0.9749 ± 0.0091) and Logistic Regression (0.9565 ± 0.0084). The low variability observed for Random Forest indicates highly stable learning behavior across different data partitions. In contrast, Naive Bayes exhibited higher variance, suggesting greater sensitivity to dataset distribution, whereas k-NN consistently performed worse due to limitations in handling high-dimensional, sparse TF-IDF representations.

To further validate whether the observed differences in model performance were statistically meaningful, paired t-tests were conducted using fold-wise Macro-F1 scores obtained from stratified 5-fold cross-validation. The paired design ensures that performance comparisons were made under identical data partitions, thereby reducing variability caused by dataset sampling.

A significance level of $\alpha = 0.05$ was initially adopted, which represents the conventional threshold indicating a 5% probability of incorrectly rejecting the null hypothesis (Type I error). However, because multiple pairwise comparisons were performed among classifiers, the Bonferroni correction was applied to control the family-wise error rate. With three planned comparisons, the adjusted significance threshold became $\alpha = 0.05/3 = 0.0167$. This stricter threshold reduces the likelihood of false-positive findings arising from repeated hypothesis testing and ensures more conservative and reliable statistical conclusions.

As summarized in Table 5, all pairwise comparisons produced p-values substantially below the adjusted

significance level ($p < 0.0167$), indicating that performance differences among the evaluated models are statistically significant. Specifically, Random Forest significantly outperformed SVM ($t = 6.8905$, $p = 0.002325$) and Logistic Regression ($t = 12.0026$, $p = 0.000276$), while SVM also demonstrated statistically superior performance compared to Logistic Regression ($t = 6.6217$, $p = 0.002698$). These results confirm that the ranking in predictive performance is consistent across cross-validation folds and unlikely to be due to random variation in the training data partitions.

Table 5. Statistical significance testing results

Comparison	t	p	Cohen's d	Significant
RF vs SVM	6.8905	0.002325	3.0815	Yes
SVM vs LR	6.6217	0.002698	2.9613	Yes
RF vs LR	12.0026	0.000276	5.3677	Yes

Beyond statistical significance, effect sizes were quantified using Cohen's d to assess the magnitude of performance differences. The obtained effect sizes ranged from 2.96 to 5.37, which far exceed the conventional threshold for a large effect ($d = 0.8$). These exceptionally large effect sizes indicate that the performance improvements are not only statistically detectable but also practically substantial. In particular, the extremely large effect observed between Random Forest and Logistic Regression ($d = 5.37$) indicates a pronounced difference in predictive performance, suggesting that ensemble-based learning provides a considerable advantage for sentiment classification using TF–IDF feature representations.

The superior performance of Random Forests can be attributed to their ensemble learning mechanism, which aggregates multiple decision trees to capture nonlinear relationships and complex feature interactions in high-dimensional textual data. In contrast, while SVM achieved strong performance due to its effectiveness in handling sparse feature spaces, its linear decision boundary may limit its ability to model more complex feature dependencies captured by ensemble methods. Logistic Regression, although stable and computationally efficient, performed comparatively poorly due to its strict linear assumptions.

The combination of low standard deviation values and statistically significant differences indicates both performance stability and consistent generalization capability across folds. This demonstrates that the observed improvements are systematic rather than dataset-specific artifacts.

From a comparative perspective, Random Forest is most appropriate when predictive performance is the top priority, even at the cost of longer training times. For contexts requiring efficiency and scalability, Naïve Bayes and Logistic Regression offer more practical solutions. SVM provides a middle ground, balancing accuracy with computational feasibility. The statistical validation supports that Random Forest provides a reliably superior modeling approach for the sentiment classification task considered in this study.

These findings reflect broader trends in machine learning research, which emphasize the importance of weighing predictive accuracy against computational cost when selecting models for real-world sentiment classification [46, 47].

While quantitative evaluation metrics demonstrate strong classification performance, qualitative analysis was conducted to determine whether higher metric scores correspond to meaningful sentiment interpretation. Representative correctly classified and misclassified samples generated by the best-performing model (Random Forest) were examined, as shown in Table 6.

Table 6. Representative examples from the qualitative evaluation of random forest predictions

Text Excerpt	Actual Sentiment	Predicted Sentiment	Interpretation
"Ma'am is a tender and warm person for her students that stimulate learning among students. I am grateful for having her as our professor."	Positive	Positive	The model correctly identifies positive affective descriptors related to teaching support.
"The ability of Sir to provide a dynamic and welcoming learning atmosphere is one of their greatest attributes."	Positive	Positive	Positive engagement keywords were strongly captured by the classifier.
"For a major subject instructor, she rarely attends our class."	Negative	Negative	Absence-related complaint correctly interpreted as negative sentiment.
"Sir is attentive and very good on presenting and teaching us the subject matter, but barely states what things we need to learn and practice so that the subject he taught us will clearly maintain to our minds."	Positive	Negative	A mixed grammatical structure weakens the positive signal detection.
"Needs improvement on interactive activities."	Negative	Neutral	Suggestive feedback is interpreted as constructive rather than explicitly negative.
"Although we rarely meet in classes, Sir is very approachable and is dynamic when it comes to getting the lectures across us students."	Positive	Neutral	Informal phrasing and fragmented syntax reduce the accuracy of sentiment intensity detection.

Correct classifications indicate that the model effectively captures sentiment-bearing linguistic patterns associated with student feedback. Expressions describing positive instructor attributes, such as warmth, approachability, and interactive teaching practices, were consistently identified as positive sentiment. Similarly, statements highlighting instructional issues, such as instructor absence, were accurately classified as negative sentiment. These results suggest that the model successfully learned semantically relevant cues linked to educational experience and learner perception.

Analysis of misclassified instances provides further insight into model behavior. Several errors occurred in the feedback, including informal grammar, shortened phrases, and ambiguous linguistic structure. For example, positively intended comments written in fragmented or conversational form were occasionally interpreted as neutral or negative due

to reduced lexical clarity. Additionally, phrases such as "need improvement" were sometimes classified as neutral, suggesting that the model may interpret advisory language as moderate sentiment rather than explicit dissatisfaction.

These findings highlight a common challenge in sentiment analysis of educational feedback: opinions are often expressed in mixed or nuanced language rather than strictly polarized sentiment. Importantly, the observed misclassifications are semantically understandable rather than arbitrary, suggesting that model errors arise from linguistic ambiguity rather than systematic bias.

The qualitative assessment confirms that Random Forest's strong quantitative performance corresponds to a meaningful interpretation of student sentiments. The model demonstrates the ability to capture contextually relevant evaluative expressions but exhibits predictable limitations when handling implicit or mixed-sentiment statements. This alignment between numerical performance and semantic interpretation supports the practical applicability of the proposed approach for analyzing educational feedback data.

Although Transformer-based and deep learning models represent the current state of the art in NLP and sentiment analysis, their evaluation falls outside the scope of this study. The objective is not to achieve maximal predictive performance, but to provide a comparative baseline of classical models that remain widely used in educational institutions due to their efficiency, interpretability, and ease of deployment.

V. Conclusion

This work conducted a comparative study of the Logistic Regression, Support Vector Machine, Naïve Bayes, Random Forest, and k-Nearest Neighbors machine learning algorithms for sentiment classification. Algorithms were compared against three main goals: assessment of predictive power, exploration of confusion matrices to assess classification validity, and a balance between accuracy and computational efficiency.

Overall, the results indicated that Random Forest achieved the highest accuracy (99%) and improved classification performance across all sentiment classes, substantiating its capacity to handle complex feature interactions. But this improved predictive capacity came at the expense of computational efficiency, with Random Forest demonstrating the longest training time (4.575s). Naïve Bayes occupied minimum time (0.019s) but slightly lower accuracy (93%). SVM was a trade-off solution, achieving maximum accuracy (97%) at a fairly good execution time. Logistic Regression also achieved competitive performance (94% accuracy), while k-NN performed worst (88% accuracy) and was thus less suitable for large-scale sentiment analysis applications.

Confusion matrix analysis further revealed that SVM and Random Forest performed better at maintaining consistency in minimizing misclassification across sentiment categories, particularly in discriminating subtle differences between positive and neutral sentiments. These findings indicate the need to balance classification performance and computational time, particularly in time-sensitive applications such as customer service tracking, financial sentiment analysis, or real-time opinion mining on social media.

Thus, this work stresses that a "best" model for sentiment classification is a function of a trade-off between accuracy and computational complexity, suited to a given application's requirements. By showing classical and state-of-the-art methods in parallel, researchers and practitioners can further tailor the strength and flexibility of sentiment analysis systems.

Based on the results, several recommendations can be proposed: (1) Practitioners aiming for maximum predictive accuracy may prioritize Random Forest, especially in offline or batch-processing environments where computational time is less restrictive. Conversely, Naïve Bayes and SVM are recommended for real-time sentiment classification tasks, where faster execution is critical, (2) Although this study focused on classical machine learning models, extending the comparative framework to include deep learning architectures (e.g., LSTMs, Transformers, BERT) would provide valuable insights into whether modern neural approaches significantly outperform traditional models in both performance and efficiency, and (3) For large-scale social media or streaming data analysis, models like SVM and Logistic Regression strike a practical balance between scalability and accuracy, making them suitable for production-level deployment. Future research may enhance the originality and explanatory power of this work by incorporating feature importance analysis using explainable AI techniques such as SHAP or LIME, enabling word-level interpretability of sentiment predictions. Additionally, extending the dataset to include cross-institutional sources and benchmarking hybrid ML-DL models may further strengthen generalizability and analytical depth.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Aaron Paul M. Dela Rosa wrote the results and discussion, and conclusion sections of the paper. He coded the Python Notebook's program to train and test the models for comparison. Renato L. Adriano II preprocessed the data and wrote the methods section. Lilibeth G. Antonio wrote the introduction and related work section. Aaron Paul M. Dela Rosa finalized the paper, and all authors approved of the final version for publication.

Acknowledgment

The authors would like to thank Mr. Hans Gabrielle B. Santos for his help in annotating the dataset.

References

[1] T. Shaik, X. H. Tao, C. Dann, H. R. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Nat. Lang. Process. J.*, vol. 2, 100003, 2023. doi: 10.1016/j.nlp.2022.100003

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014. doi: 10.1016/j.asej.2014.04.011

[3] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst*., vol. 28, no. 2, pp. 15–21, 2013. doi: 10.1109/MIS.2013.30

[4] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev.: Data Mining Knowl. Discov*., vol. 8, no. 4, e1253, 2018. doi: 10.1002/widm.1253

[5] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Sentiment analysis: Towards a tool for analyzing real-time students feedback," in *Proc. IEEE 26th Int. Conf. Tools Artif. Intell.*, 2014, pp. 419–423. doi: 10.1109/ICTAI.2014.70

[6] R. S. Baragash, H. Aldowah, and I. N. Umar, "Students' perceptions of e-learning in Malaysian universities: Sentiment analysis based machine learning approach," *J. Inf. Technol. Educ.: Res*., vol. 21, pp. 439–463, 2022. doi: 10.28945/5024

[7] S. K. D'Mello and A. C. Graesser, "Feeling, thinking, and computing with affect-aware learning technologies," in *The Oxford Handbook of Affective Computing*, R. A. Calvo, S. K. D'Mello, J. Gratch, and A. Kappas, Eds., Oxford, U.K.: Oxford Univ. Press, 2015, pp. 419–434. doi: 10.1093/oxfordhb/9780199942237.013.032

[8] X. H. Tao, A. Shannon-Honson, P. Delaney, C. Dann, H. R. Xie, Y. Li, and S. O'Neill, "Towards an understanding of the engagement and emotional behaviour of MOOC students using sentiment and semantic features," *Comput. Educ.: Artif. Intell*., vol. 4, 100116, 2023. doi: 10.1016/j.caeai.2022.100116

[9] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev*., vol. 53, pp. 4335–4385, 2020. doi: 10.1007/s10462-019-09794-5

[10] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, 2019. doi: 10.1016/j.eswa.2018.10.003

[11] M. R. Baker and A. Utku, "Unraveling user perceptions and biases: A comparative study of ML and DL models for exploring twitter sentiments towards ChatGPT," *J. Eng. Res*., vol. 13, no. 2, pp. 1658–1665, 2025. doi: 10.1016/j.jer.2023.11.023

[12] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decis. Anal. J.*, vol. 11, 100473, 2024. doi: 10.1016/j.dajour.2024.100473

[13] J. Y. Tan, A. C. S. Kit, and C. W. Tan, "A comparative study of machine learning algorithms for sentiment analysis of game reviews," *J. Inst. Eng. Malaysia*, vol. 82, no. 3, 2022. doi: 10.54552/v82i3.101

[14] P. Shah, H. Patel, and P. Swaminarayan, "Multitask sentiment analysis and topic classification using BERT," *EAI Endorsed Trans. Scalable Inf. Syst*., vol. 12, no. 1, 2024. doi: 10.4108/eetsis.5287

[15] Al-Saqqa, N. Obeid, and A. Awajan, "Sentiment analysis for Arabic text using ensemble learning," in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, 2018, pp. 1–7. doi: 10.1109/AICCSA.2018.8612804

[16] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for test classification using feature ranking and optimal number of trees," *J. King Saud Univ. – Comput. Inf. Sci*., vol. 34, no. 6, pp. 2733–2742, 2022. doi: 10.1016/j.jksuci.2022.03.012

[17] Z. Kastrati, F. Dalipi, A. S. Imra, K. P. Nuci, and M. Ahmad Wani, "Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study," *Appl. Sci*., vol. 11, no. 9, 3986, 2021. doi: 10.3390/app11093986

[18] A. Abdi, G. Sedrakyan, B. Veldkamp, J. V. Hillegersberg, and S. M. van den Berg, "Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems," *Soft Comput*., vol. 27, pp. 14073–14094, 2023. doi: 10.1007/s00500-023-07926-2

[19] I. Jahan, M. N. Islam, M. M. Hasan, and M. R. Siddiky, "Comparative analysis of machine learning algorithms for sentiment classification in social media text," *World J. Adv. Res. Rev.*, vol. 23, no. 3, pp. 2842–2852, 2024. doi: 10.30574/wjarr.2024.23.3.2983

[20] S. A. Patil, K. P. Adhiya, and G. K. Patnaik, "Sentiment analysis of student feedback using lexicon based method and hybrid machine learning method," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 12s, pp. 137–145, 2024.

[21] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Comput. Surv*., vol. 54, no. 3, pp. 1–40, 2021. doi: 10.1145/3439726

[22] K. Walji, A. Erraissi, A. Zakrani, and M. Banane, "Comparative performance evaluation of machine learning algorithms in sentiment analysis," in *Proc. Int. Conf. Artif. Intell. Smart Environ*., 2025, pp. 121–128. doi: 10.1007/978-3-031-90921-4_17

[23] F. Wang, J. L. Zhang, Y. C. Li, K. Deng, and J. S. Liu, "Bayesian text classification and summarization via a class-specified topic model," *J. Mach. Learn. Res*., vol. 22, pp. 1–48, 2021.

[24] R. K. Das, M. Islam, M. M. Hasan, S. Razia, M. Hassan, and S. A. Khushbu, "Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models," *Heliyon*, vol. 9, no. 9, e20281, 2023. doi: 10.1016/j.heliyon.2023.e20281

[25] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, "Sentiment analysis using SVM: A systematic literature review," *Int. J. Adv. Comput. Sci. Appl*., vol. 9, no. 2, 2018. doi: 10.14569/IJACSA.2018.090226

[26] Y. Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. – Comput. Inf. Sci*., vol. 36, no. 4, 102048, 2024. doi: 10.1016/j.jksuci.2024.102048

[27] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, 150, 2019. doi: 10.3390/info10040150

[28] L. Ashbaugh and Y. Zhang, "A comparative study of sentiment analysis on customer reviews using machine learning and deep learning," *Computers*, vol. 13, no. 12, 340, 2024. doi: 10.3390/computers13120340

[29] B. Bahrawi, "Sentiment analysis using Random Forest algorithm-online social media based," *J. Inf. Technol. Utilization*, vol. 2, no. 2, pp. 29–33, 2019. doi: 10.30818/jitu.2.2.2695

[30] H. Elfaik and E. H. Nfaoui, "A comparative analysis of classification algorithms for sentiment analysis using word embeddings," *Adv. Intell. Syst. Sustain. Dev. (AI2SD' 2019)*, Springer, 2020, pp. 1–11. doi: 10.1007/978-3-030-36674-2_1

[31] C. J. Wang, F. X. You, and Y. Wang, "Text intelligent classification model based on improved KNN algorithm in library service transformation," *Syst. Soft Comput*., vol. 7, 200313, 2025. doi: 10.1016/j.sasc.2025.200313

[32] N. AlGhamdi and S. Khatoon, "Improving sentiment prediction using heterogeneous and homogeneous ensemble methods: A comparative study," *Procedia Comput. Sci*., vol. 194, pp. 60–68, 2021. doi: 10.1016/j.procs.2021.10.059

[33] S. Modha, P. Majumder, and T. Mandl, "An empirical evaluation of text representation schemes to filter the social media stream," *J. Exp. Theor. Artif. Intell.*, vol. 34, no. 3, pp. 499–525, 2022. doi: 10.1080/0952813X.2021.1907792

[34] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models," in *Proc. 1st Workshop Eval. Comparison NLP Syst. (Eval4NLP)*, 2020, pp. 79–91.

[35] S. Nawaz, "A comparative study of machine learning algorithms for sentiment analysis," *Int. J. Res. Publ. Rev*., vol. 4, no. 8, pp. 199–2003, 2023.

[36] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci*., vol. 5, no. 7, e623, 2021. doi: 10.7717/peerj-cs.623

[37] Y. K. Li, "Review on natural language processing models," *Appl. Comput. Eng*., vol. 35, no. 1, pp. 1–7, 2024. doi: 10.54254/2755-2721/35/20230350

[38] C. G. Ozmen and S. Gunduz, "Comparison of machine learning models for sentiment analysis of big Turkish web-based data," *Appl. Sci*., vol. 15, no. 5, 2297, 2025. doi: 10.3390/app15052297

[39] O. Rainio, J. Teuho, and R. Klen, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep*., vol. 14, p. 6086, 2024. doi: 10.1038/s41598-024-56706-x

[40] G. Canbek, T. T. Temizal, and Sagiroglu, "PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics," *SN Comput. Sci*., vol. 4, p. 13, 2023. doi: 10.1007/s42979-022-01409-1

[41] A. Al Fahoum and T. A. Ghobo, "Performance predictions of sci-fi films via machine learning," *Appl. Sci*., vol. 13, no. 7, 4312, 2023. doi: 10.3390/app13074312

[42] Z. Khanam, "Sentiment analysis of user reviews in an online learning environment: Analyzing the methods and future prospects," *Eur. J. Educ. Pedagogy*, vol. 4, no. 2, pp. 209–217, 2023. doi: 10.24018/ejedu.2023.4.2.531

[43] M. Cherradi and A. El Haddadi, "Comparative analysis of machine learning algorithms for sentiment analysis in film reviews," *Acadlore Trans. AI Mach. Learn*., vol. 3, no. 3, pp. 137–147, 2024. doi: 10.56578/ataiml030301

[44] K. Li, Y. Xie, S. Li, Z. Liu, and D. Liu, "Research on sentiment analysis based on ensemble learning," *Int. J. Comput. Sci. Inf. Technol.*, vol. 4, no. 3, pp. 160–170, 2024. doi: 10.62051/ijcsit.v4n3.17

[45] F. Panjaitan, W. Ce, H. Oktafiandi, G. Kanugrahan, Y. Ramdhani, and V. H. C. Putra, "Evaluation of machine learning models for sentiment analysis in the South Sumatra governor election using data balancing techniques," *J. Inf. Syst. Inform*., vol. 7, no. 1, pp. 461–478, 2025. doi: 10.51519/journalisi.v7i1.1019

[46] M. Alzaid and F. Fkih, "Sentiment analysis of students' feedback on e-learning using a hybrid fuzzy model," *Appl. Sci*., vol. 13, no. 23, 12956, 2023. doi: 10.3390/app132312956

[47] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Expert Syst. Appl.*, vol. 201, 117193, 2022. doi: 10.1016/j.eswa.2022.117193