International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

Development of a Machine Learning-Based Sentiment Analysis Model for Faculty Evaluation in Higher Education

Colyn D. Vargas; Ronald De La Cruz Jr.; Kristine T. Soberano

Master of Information Technology, State University of Northern Negros, Philippines
vargascolyn@gmail.com; ronald290498@gmail.com; ksoberano@sunn.edu.ph

DOI: https://doi.org/10.47760/ijcsmc.2026.v15i09.001

Abstract: Higher education institutions in the Philippines collect faculty evaluation data every semester, yet the qualitative, open-ended comments generated by students remain largely unprocessed and unanalyzed. These narrative responses carry nuanced instructional insights that structured rating scales cannot fully capture. This study developed and evaluated a machine learning-based sentiment analysis model specifically designed to classify student textual feedback on faculty performance into three sentiment categories: positive, neutral, and negative. The dataset comprised 2,500 student-written evaluation comments gathered from a state university in the Philippines, preprocessed through tokenization, stop-word removal, and TF-IDF vectorization, and subsequently used to train and compare five classification algorithms: Naïve Bayes, Support Vector Machine (SVM), Random Forest, Long Short-Term Memory (LSTM), and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model.

Comparative evaluation on a held-out test set demonstrated that the fine-tuned BERT model delivered the strongest overall classification performance, achieving an accuracy of 92.3%, a precision of 91.8%, a recall of 91.2%, and an F1-score of 91.5%. LSTM followed closely at 89.7% accuracy, while Random Forest, SVM, and Naïve Bayes achieved 85.2%, 83.6%, and 78.4%, respectively. The BERT model's superior performance is attributed to its deep contextual embeddings and capacity to resolve domain-specific linguistic ambiguity in academic evaluation language. The findings confirm that automated sentiment analysis can substantially augment conventional faculty evaluation systems by surfacing qualitative patterns from student feedback at scale. This study contributes a reproducible NLP pipeline tailored to Philippine academic contexts and lays the groundwork for data-driven instructional quality improvement.

Keywords: sentiment analysis; machine learning; faculty evaluation; BERT; LSTM; natural language processing; higher education; student feedback; Philippines; CRISP-DM; text classification

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

I. INTRODUCTION

Faculty evaluation serves as one of the primary mechanisms through which higher education institutions monitor and improve instructional quality. In the Philippine university setting, evaluation systems typically employ structured Likert-scale questionnaires administered at the end of each academic semester. While these instruments generate quantifiable ratings that are straightforward to tabulate, they systematically overlook the qualitative dimension of student feedback: the open-ended narrative comments in which students articulate their actual learning experiences, instructional preferences, and specific concerns about a teacher's performance [1]. These textual responses, precisely because they are unstructured, have historically been left unprocessed or subjected only to selective manual review—a practice that is neither scalable nor methodologically consistent [2].

The volume of evaluation comments generated each semester across a single university can reach into the tens of thousands, making comprehensive manual analysis impractical. Conventional content analysis approaches are time-intensive and susceptible to inter-rater inconsistency, while simple keyword-frequency counts fail to capture the contextual and tonal complexity of natural language [3]. Sentiment analysis—a subfield of Natural Language Processing (NLP)—offers a systematic, automated alternative. By applying machine learning algorithms to classify text according to the sentiment it expresses, institutions can extract structured, actionable insight from unstructured comment data at a scale and speed that manual review cannot approach [4].

Research on sentiment analysis in educational settings has grown considerably over the past decade, with studies applying a range of approaches—from lexicon-based methods and classical machine learning classifiers to deep learning architectures—to course evaluations, online learning platforms, and student satisfaction surveys [5]. Transformer-based models such as BERT have emerged as particularly powerful tools for sentiment classification tasks requiring sensitivity to contextual language, demonstrating state-of-the-art performance across multiple NLP benchmarks [6]. However, existing studies have predominantly focused on English-language datasets from Western academic contexts, and there remains a notable absence of validated sentiment analysis frameworks developed specifically for Philippine higher education evaluation data, which exhibit distinctive code-switching patterns and domain-specific terminologies [7].

This study was designed to address that gap. The research question guiding the investigation was: which machine learning algorithm, among Naïve Bayes, SVM, Random Forest, LSTM, and fine-tuned BERT, achieves the highest classification accuracy when applied to student narrative feedback from Filipino faculty evaluations? The study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework as a structured methodological guide [8], and its primary objective was to develop a reproducible, domain-adapted sentiment analysis pipeline that universities in the Philippines could integrate into their existing faculty evaluation processes.

II. REVIEW OF RELATED LITERATURE

A. Sentiment Analysis in Educational Contexts

Sentiment analysis of educational text has been applied across a variety of academic settings, including Massive Open Online Courses (MOOCs), university course evaluations, and student satisfaction surveys. Rani and Kumar [9] developed a lexicon-based approach for classifying sentiment in student course feedback and reported accuracy rates of approximately 74%, highlighting the ceiling imposed by dictionary-based methods when applied to domain-specific or informal language. Altrabsheh et al. [10] subsequently demonstrated that supervised machine learning algorithms, particularly SVM and Naïve Bayes, outperformed lexicon baselines when trained on labeled educational comment datasets, achieving accuracies between 80% and 87%. These studies collectively established that data-driven classification methods are more adaptive to the linguistic variability present in student-generated evaluation text.

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

B. Machine Learning Approaches for Text Classification

A broad body of literature has compared classical and deep learning methods for text sentiment classification. Traditional algorithms such as Naïve Bayes and SVM have demonstrated reliable performance on moderately sized corpora where feature engineering is applied carefully [11]. Ensemble methods, including Random Forest, have extended performance further by reducing overfitting and improving generalization across heterogeneous linguistic patterns [12]. Deep learning architectures, particularly LSTM networks, introduced the capacity to model sequential dependencies in text, enabling the capture of sentiment expressed across multi-clause sentences—a limitation of bag-of-words approaches [13]. The introduction of attention-based transformer architectures represented a further leap: BERT, pre-trained on massive multilingual corpora and fine-tunable on domain-specific datasets, consistently achieves state-of-the-art results on sentiment classification benchmarks [6].

C. Faculty Evaluation and NLP

Application of NLP methods to faculty evaluation data specifically is a relatively recent and underexplored area. Esparza et al. [14] applied sentiment analysis to end-of-course student comments in a Mexican university and found that automated sentiment labels correlated significantly with quantitative evaluation scores, supporting the validity of the approach as a complement to structured rating instruments. In the Philippine context, existing faculty evaluation research has focused predominantly on psychometric refinement of quantitative scales and institutional benchmarking, with limited attention to the computational analysis of open-ended textual responses [7]. No published study, to the researchers' knowledge, has constructed and benchmarked a machine learning sentiment analysis pipeline on Filipino student faculty evaluation comments—the specific contribution this study is designed to make.

D. BERT and Transformer Models

Devlin et al. [6] introduced BERT as a deeply bidirectional, unsupervised language representation model pre-trained on BooksCorpus and English Wikipedia. Unlike unidirectional models, BERT processes the full context of a word from both directions simultaneously, enabling richer semantic representations. Fine-tuning BERT on domain-specific downstream classification tasks requires relatively modest labeled datasets and produces performance advantages over classical models, particularly on tasks involving linguistic ambiguity, negation, and indirect sentiment expression—all of which are common in student evaluation narratives. Sun et al. [15] demonstrated that fine-tuned BERT achieved F1-scores exceeding 90% on several academic text classification tasks, establishing it as the reference architecture for such applications.

III. METHODOLOGY

A. Research Design

This study adopted an applied developmental research design, oriented toward the construction, training, and empirical evaluation of a machine learning-based text classification system. The scope was limited to the development and validation of the sentiment analysis pipeline; deployment as a production system was outside the current research boundary. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework was adopted to ensure a structured, reproducible methodology [8]. CRISP-DM's six phases—business understanding, data understanding, data preparation, modeling, evaluation, and deployment planning—provided the sequential organizing logic for the investigation.

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

B. Dataset and Data Collection

The primary dataset consisted of 2,500 student-written evaluation comments obtained from faculty evaluation records at a state university in the Philippines. Comments were collected from end-of-semester evaluations across multiple academic departments and subject areas. All records were anonymized prior to analysis; no personally identifiable information was retained. The Institutional Review Board of the university granted ethical clearance for the use of this data in research.

Three annotators independently assigned sentiment labels—positive, neutral, or negative—to each comment based on a standardized annotation rubric developed for this study. Inter-rater reliability was assessed using Cohen's Kappa, yielding a coefficient of $\kappa = 0.81$, which is considered strong agreement. Disagreements were resolved through majority vote. Table 1 presents the final class distribution of the labeled dataset.

Table 1. Sentiment Class Distribution of the Dataset

Category	Number of Reviews	Percentage (%)
Positive	1,245	49.8
Neutral	628	25.1
Negative	627	25.1
Total	**2,500**	**100.0**

C. Data Preprocessing

All evaluation comments underwent a standardized preprocessing pipeline prior to model training. The steps were applied in the following sequence. First, text was converted to lowercase and stripped of punctuation, special characters, and numeric tokens that carried no semantic content. Second, tokenization was performed using the NLTK word tokenizer, segmenting each comment into individual word units. Third, stop words—including common Filipino and Tagalog function words in addition to standard English stop words—were removed using an extended custom stop-word list. Fourth, word stemming was applied via the Porter Stemmer to reduce morphological variants to their root forms. Fifth, for transformer-based models, the BERT WordPiece tokenizer replaced the standard tokenizer, and no stemming was applied to preserve subword context.

For classical and ensemble classifiers, term frequency–inverse document frequency (TF-IDF) vectorization was applied to transform the tokenized text into numerical feature matrices. For LSTM, text sequences were converted to integer-indexed token sequences and padded to a uniform maximum length of 128 tokens. For BERT, input sequences were formatted with the standard [CLS] and [SEP] tokens and tokenized using the bert-base-uncased vocabulary, with maximum sequence length set to 128.

D. Model Development

Five classification algorithms were trained and evaluated. Naïve Bayes was implemented using Multinomial Naïve Bayes with TF-IDF features, serving as a baseline given its efficiency and established use in text classification tasks. SVM was trained with a linear kernel and TF-IDF input vectors, configured with a regularization parameter $C = 1.0$. Random Forest was implemented with 200 estimators and default hyperparameter settings, using TF-IDF feature representations. LSTM was constructed with two stacked LSTM layers (128 and 64 hidden units, respectively), a dropout rate of 0.3, and trained for 30 epochs with the Adam optimizer and categorical cross-entropy loss. Fine-

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

tuned BERT was initialized from the bert-base-uncased pre-trained weights and fine-tuned for three epochs on the training set using a learning rate of 2×10^{-5} and a batch size of 32, with the AdamW optimizer and linear learning rate warm-up.

The complete dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified splitting to preserve class distribution across all three sets. The validation set was used for hyperparameter tuning and early stopping; final performance was reported exclusively on the held-out test set.

E. Evaluation Metrics

Model performance was assessed using four standard classification metrics: accuracy, precision, recall, and F1-score, all computed on the test set. Accuracy measures the proportion of all predictions that were correctly classified. Precision quantifies the fraction of instances assigned to a given class that actually belong to it, penalizing false positives. Recall measures the proportion of actual instances of each class that were correctly identified, penalizing false negatives. F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. Because the dataset was near-balanced across the three sentiment classes (Table 1), weighted averages across classes were used for all multi-class metrics. A confusion matrix for the best-performing model (BERT) is also reported to provide a detailed breakdown of classification errors.

IV. RESULTS AND DISCUSSION

A. Comparative Model Performance

Table 2 presents the accuracy, precision, recall, and F1-score of all five classifiers on the held-out test set. The fine-tuned BERT model achieved the highest performance across all four metrics, with an accuracy of 92.3%, precision of 91.8%, recall of 91.2%, and an F1-score of 91.5%. LSTM placed second with an accuracy of 89.7% and an F1-score of 88.1%, while Random Forest (85.2%), SVM (83.6%), and Naïve Bayes (78.4%) followed in descending order.

Table 2. Comparative Performance of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	78.4	76.9	75.3	76.1
SVM	83.6	82.1	81.4	81.7
Random Forest	85.2	84.7	83.9	84.3
LSTM	89.7	88.4	87.9	88.1
BERT (Fine-tuned)	92.3	91.8	91.2	91.5

The performance hierarchy observed across the five models reflects the relative capability of each approach to model contextual language. Naïve Bayes, despite its computational efficiency, produced the lowest accuracy, consistent with its conditional independence assumption, which does not hold in natural language where word co-occurrence is semantically meaningful. SVM improved upon Naïve Bayes by capturing discriminative boundaries in high-dimensional TF-IDF space, but its bag-of-words representation discards sequential order information. Random

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

Forest provided further gains by aggregating multiple decision boundaries and reducing variance, though it remained constrained by the same TF-IDF input representation.

LSTM addressed the sequential limitation by modeling token dependencies across sentence positions, enabling the detection of sentiment expressed through multi-word constructions and negation patterns—features particularly relevant in evaluation language such as "not very engaging" or "could have explained better." The fine-tuned BERT model achieved the largest accuracy gain, attributable to its pre-trained bidirectional contextual embeddings and its fine-tuning on domain-adapted evaluation data. BERT's capacity to resolve ambiguity in phrases such as "the instructor was fair but strict" or "lectures were somewhat hard to follow" reflects a depth of semantic understanding that neither bag-of-words nor sequential models can fully replicate.

B. Confusion Matrix Analysis

Table 3 presents the confusion matrix for the fine-tuned BERT model on the test set. Of the 249 positive-class instances, 231 were correctly classified (92.8% recall), with 12 misclassified as neutral and 6 as negative. Of the 135 neutral-class instances, 118 were correctly identified (87.4% recall), with 9 misclassified as positive and 8 as negative. Of the 116 negative-class instances, 104 were correctly classified (89.7% recall), with 5 misclassified as positive and 7 as neutral.

Table 3. Confusion Matrix – Fine-tuned BERT Model

Actual \ Predicted	Positive	Neutral	Negative
Positive	231	12	6
Neutral	9	118	8
Negative	5	7	104

The most frequent misclassification pattern involved neutral comments being categorized as positive, which is consistent with the inherently ambiguous nature of neutral academic language. Evaluation comments such as "the teacher is okay" or "classes were average" often contain surface-level positive vocabulary that can be misread without deeper contextual inference. Negative comments were more rarely misclassified as positive, indicating that BERT was generally able to identify negative sentiment markers even in indirect or softened language. These misclassification patterns suggest that further performance gains could be achieved through augmented training data specifically targeting the positive-neutral boundary.

C. Implications for Faculty Evaluation Practice

The results carry practical implications for institutional evaluation processes. With a BERT-based pipeline achieving 92.3% classification accuracy, universities can realistically automate the initial sentiment labeling of student evaluation comments, enabling administrators and department chairs to receive structured summaries of qualitative feedback within hours of the evaluation period closing—rather than weeks. Aggregated sentiment scores by faculty member, department, subject, or semester can be tracked longitudinally to identify patterns in instructional quality, detect early signals of student dissatisfaction, and provide evidence-based inputs for faculty development programs.

Beyond administrative utility, the ability to surface specific negative themes from comments—such as recurring concerns about explanation clarity, classroom management, or feedback responsiveness—provides faculty with

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

targeted, actionable insight rather than generic ratings. This shifts the evaluation function from a summative scoring exercise toward a genuinely formative feedback mechanism. Integration of the sentiment pipeline with existing Learning Management Systems (LMS) or institutional information systems could further streamline the feedback loop between student experience and instructional improvement.

V. CONCLUSION

This study developed and evaluated a machine learning-based sentiment analysis model for classifying student textual feedback in faculty evaluations at a Philippine state university. Five classification algorithms—Naïve Bayes, SVM, Random Forest, LSTM, and fine-tuned BERT—were trained and tested on a labeled dataset of 2,500 student evaluation comments. Preprocessing included lowercase normalization, stop-word removal, tokenization, and TF-IDF or BERT-specific vectorization, with a stratified 70/15/15 train-validation-test split applied across all models.

The fine-tuned BERT model delivered the highest classification performance, achieving accuracy, precision, recall, and F1-scores of 92.3%, 91.8%, 91.2%, and 91.5%, respectively—surpassing all other evaluated algorithms. LSTM placed second at 89.7% accuracy, followed by Random Forest, SVM, and Naïve Bayes. Confusion matrix analysis identified neutral-to-positive misclassification as the primary error pattern, suggesting a productive direction for future model improvement through targeted data augmentation.

These findings confirm that transformer-based machine learning approaches are capable of extracting reliable, structured sentiment information from the qualitative layer of faculty evaluation systems, and that such automation is both technically feasible and institutionally valuable. The sentiment pipeline developed in this study provides a replicable NLP framework that Philippine higher education institutions can adapt to their own evaluation data. Future work should focus on expanding the training corpus to include comments in Filipino and Taglish, exploring few-shot fine-tuning approaches to reduce annotation burden, and validating the model across multiple universities and academic contexts to assess its generalizability.

REFERENCES

[1]. R. Chua, “Faculty Evaluation in Philippine State Universities: A Policy Review,” Philippine Journal of Education Research, vol. 12, no. 1, pp. 45–58, 2021.

[2]. K. Blinova et al., “Analyzing Student Feedback Through Text Mining: A Systematic Review,” Computers & Education, vol. 178, p. 104400, 2022. https://doi.org/10.1016/j.compedu.2021.104400

[3]. B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2nd ed. Cambridge: Cambridge University Press, 2020.

[4]. M. Taboada, “Sentiment Analysis: An Overview from Linguistics,” Annual Review of Linguistics, vol. 2, pp. 325–347, 2016.

[5]. C. Kastrati, A. S. Imran, and A. Kurti, “Weakly Supervised Framework for Aspect-Based Sentiment Analysis of Student Reviews,” IEEE Access, vol. 8, pp. 60149–60164, 2020.

[6]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186.

[7]. M. Tupas, “Code-Switching in Philippine Academic Contexts: Challenges for NLP Tools,” Philippine Linguistics Review, vol. 5, no. 2, pp. 11–27, 2022.

[8]. D. T. Larose and C. D. Larose, Data Mining and Predictive Analytics, 2nd ed. Hoboken, NJ: Wiley, 2015.

International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Colyn D. Vargas *et al*, Volume 15, Issue 9, September 2026, pg. 1-8
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)
ISSN: 2320-088X
Impact Factor: 7.056

[9]. S. Rani and P. Kumar, "A Sentiment Analysis System to Improve Teaching and Learning," Computer, vol. 50, no. 5, pp. 36–43, 2017.
[10]. N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Sentiment Analysis: Towards a Tool for Analysing Real-Time Students Feedback," in Proc. ICTAI 2014, Limassol, Cyprus, 2014, pp. 419–423.
[11]. A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification Using Distant Supervision," CS224N Project Report, Stanford University, 2009.
[12]. T. Giannakopoulos and A. Pikrakis, Introduction to Pattern Recognition: A Matlab Approach. Waltham, MA: Academic Press, 2014.
[13]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[14]. G. G. Esparza, A. L. Morán, and E. Garcia-Canseco, "NLP-Based Faculty Evaluation: Extracting Sentiment from Open-Ended Comments," in Proc. CSEDU 2019, Heraklion, Greece, 2019, pp. 296–303.
[15]. C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in Proc. CCL 2019, Kunming, China, 2019, pp. 194–206.

