education sciences

MDPI

Article

A Context-Aware Framework for Sentiment Analysis of Student Feedback to Inform Educational Strategies in Latin America

Anabel Pineda-Briseño [1,*], Jimy Oblitas Cruz [2], Laura Cleofas Sánchez [3], Wendy Sanchez [1] and Rosario Baltazar [4]

[1] Instituto Tecnológico de Matamoros, Tecnológico Nacional de México, Matamoros 87490, Mexico; wendy.sg@matamoros.tecnm.mx
[2] Facultad de Ingeniería, Universidad Privada del Norte, Campus Cajamarca, Cajamarca 06002, Peru; jimy.oblitas@upn.edu.pe
[3] Tecnológico de Estudios Superiores de Tianguistenco, Tecnológico Nacional de México, Santiago Tianguistenco 52650, Mexico; laura_cs@test.edu.mx
[4] Instituto Tecnológico de León, Tecnológico Nacional de México, León 37290, Mexico; rosario.baltazar@leon.tecnm.mx
* Correspondence: anabel.pb@matamoros.tecnm.mx

Abstract

Understanding student feedback is essential for informing pedagogical strategies and institutional decision-making in higher education. Sentiment analysis offers scalable mechanisms for extracting insights from open-ended student evaluations; however, many existing approaches prioritize technical performance without sufficient consideration of contextual and institutional constraints, particularly in underrepresented regions. This study proposes a context-aware framework for sentiment analysis of student feedback, designed to support educational decision-making within Latin American universities. Rather than introducing new algorithms, the framework systematically evaluates established machine learning and deep learning models through a multi-phase process that includes data preprocessing, Bayesian optimization, threshold calibration, and class balancing. The framework is validated using authentic Spanish-language student feedback collected from a public university in Peru. Experimental results indicate that while advanced models can achieve strong predictive performance, simpler and more interpretable approaches often provide comparable institutional value when deployment feasibility, computational efficiency, and transparency are considered. These findings highlight that marginal performance gains do not necessarily translate into meaningful advantages for routine educational use. Overall, this work contributes a replicable and resource-sensitive framework that bridges learning analytics research and practical educational application. By prioritizing contextual suitability and interpretability, the proposed approach enables higher education institutions to leverage student sentiment data as an actionable input for continuous improvement and evidence-based educational strategies.

Keywords: sentiment analysis; student feedback; machine learning in education; educational quality assurance; data-driven decision making; Latin American higher education

Check for updates

Academic Editors: Janet Clinton and Jerome D'Agostino

Received: 26 October 2025
Revised: 3 February 2026
Accepted: 10 February 2026
Published: 5 March 2026

1. Introduction

In recent years, sentiment analysis has emerged as a powerful tool for understanding student perceptions and enhancing educational quality (Grimalt-Álvaro & Usart, 2024). By applying natural language processing (NLP) techniques to open-ended feedback, institutions can uncover valuable insights into learner engagement, instructional effectiveness, and areas requiring pedagogical improvement (Brânzilă, 2024; Fatima et al., 2023). As higher education becomes increasingly data-driven, the integration of artificial intelligence (AI) into academic decision-making has opened new pathways to personalize learning and optimize support services (Toçoğlu & Onan, 2021).

In higher education institutions, student feedback plays a critical role in improving teaching quality, curriculum design, and institutional decision-making. However, as universities increasingly rely on large-scale, open-ended feedback mechanisms, the resulting volume and heterogeneity of student comments exceed the practical limits of manual analysis, constraining institutions' capacity to extract timely and actionable insights. Manual qualitative review, while rich in insight, is often impractical at scale and limits the capacity of institutions to respond proactively to emerging concerns, particularly in resource-constrained contexts.

Recent research has applied sentiment and content analysis to teacher-narratives in higher education contexts (Barbosa & Marín-Suelves, 2024) and noted the growing role of learning analytics dashboards to interpret student data streams (Masiello et al., 2024).

Prior studies have demonstrated the utility of sentiment classification in evaluating student attitudes toward courses, faculty performance, and institutional resources (Alamoudi & Alghamdi, 2021; Onan, 2021). Traditional machine learning models such as Logistic Regression and Support Vector Machines are frequently adopted due to their interpretability and computational efficiency (Abubakar et al., 2022). More recently, ensemble approaches like Random Forest and XGBoost, as well as deep learning architectures such as Long Short-Term Memory (LSTM) networks, have shown improved capacity to capture sentiment polarity in unstructured student narratives (Iparraguirre-Villanueva et al., 2023; Reddy et al., 2022; Shen et al., 2020).

A particularly relevant example is the work of Dake and Gyimah (2023), who applied sentiment analysis to open-ended student responses in higher education. Their findings highlight how sentiment polarity can support formative assessment and provide instructors with actionable insights into student perceptions.

While prior studies have demonstrated the potential of sentiment analysis and natural language processing techniques for educational feedback analysis, much of the existing literature focuses either on algorithmic performance in isolation or on well-resourced institutional settings. As a result, there remains limited empirical evidence on how different modeling strategies, enhancement techniques, and evaluation criteria translate into actionable insights for higher education institutions operating under practical constraints, particularly in underrepresented regions such as Latin America.

Despite these advances, most existing research has focused on datasets from North America, Europe, or Asia, with limited exploration of Latin American educational contexts (Cechinel et al., 2020; Pereiro et al., 2022; Salas-Pilco & Yang, 2022). A recent systematic review highlights the growing but still uneven adoption of artificial intelligence technologies in Latin American higher education, particularly for learning analytics and decision support (Salas-Pilco & Yang, 2022). Such work reinforces the need for regional frameworks like the one proposed in this study. Furthermore, challenges related to model optimization, threshold calibration, and class imbalance remain underexplored in educational sentiment classification. These aspects are critical for translating predictive accuracy into actionable insights for institutional improvement (Hancock et al., 2022; Rodríguez Velasco et al., 2023).

In response to this gap, this study proposes a context-aware, multi-phase framework for the analysis of student feedback that emphasizes not only predictive performance but also interpretability, computational efficiency, and institutional applicability. Rather than introducing new algorithms, the framework systematically evaluates established models under different enhancement strategies, with the explicit goal of informing decision-making processes within higher education institutions. The proposed approach is designed to support interdisciplinary teams or institutional units responsible for quality assurance, analytics, or academic development, rather than individual instructors acting in isolation.

This study addresses these gaps by presenting a comprehensive, multi-phase framework for evaluating and optimizing sentiment analysis models based on real-world student feedback from a multi-campus university in Peru. The methodology compares six classification models—Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and LSTM—across multiple enhancement strategies, including feature engineering, Bayesian optimization, threshold adjustment, and class balancing using SMOTE (Chawla et al., 2002; Mary & Claral, 2025; Viveka & Priya, 2024).

By applying both conventional and deep learning techniques to a real-world and underrepresented educational context, this study contributes to a more inclusive and robust application of sentiment analysis in higher education. Recent research has shown that real-time sentiment analysis—whether based on facial emotion recognition (Tian et al., 2022) or NLP techniques applied to written feedback (Patil et al., 2024)—can provide meaningful insights into student engagement and teaching quality.

This study moves beyond technical optimization to embrace a broader goal: helping educational institutions—especially those in Latin America—make better, more timely decisions grounded in what students actually express in their feedback. It builds on prior work but places particular emphasis on transforming student voices into actionable insights that enhance educational quality, evidence-based decision making, and institutional accountability. The framework facilitates the adoption of more responsive and data-informed strategies by universities to enhance teaching and learning by converting open-ended feedback into structured, interpretable insights. These insights can directly contribute to the design of learning environments that reflect the actual needs and experiences of students, compliance with accreditation standards, and continuous improvement efforts. The research thus serves as a bridge between the practical, ordinary challenges that educational institutions that are dedicated to promoting educational quality and equity encounter and artificial intelligence.

While the sentiment analysis models evaluated in this study are well established in the natural language processing literature, their comparative performance has rarely been examined using authentic Spanish-language student feedback from Latin American higher education institutions. Accordingly, the primary contribution of this work does not lie in proposing novel algorithms, but in offering a validated and replicable, context-aware framework that integrates preprocessing strategies, Bayesian optimization, threshold calibration, and class balancing techniques. Beyond technical evaluation, the framework is explicitly designed to support pedagogical and strategic decision-making by translating model performance into criteria relevant for institutional adoption, particularly in underrepresented contexts with limited computational resources. Although the framework is grounded in data collected from Peruvian universities, its design and implementation principles are adaptable to other Latin American higher education institutions facing similar challenges in large-scale feedback analysis. The empirical evaluation is based on authentic student feedback collected through institutional surveys at a public university in Peru, reflecting real-world educational contexts within Latin American higher education rather than simulated or social-media-derived data.

Accordingly, this study pursues three main objectives. First, it aims to comparatively evaluate the performance of multiple sentiment analysis models applied to authentic Spanish-language student feedback within higher education contexts. Second, it examines how different enhancement strategies—such as data balancing, threshold calibration, and model optimization—affect both predictive outcomes and practical feasibility. Third, it seeks to translate technical evaluation results into actionable criteria that can guide institutional decision-making, particularly in contexts where interpretability, resource constraints, and scalability are central considerations.

2. Related Work

Sentiment analysis in education has become a valuable approach to assess student engagement and improve learning outcomes. A broader review of AI applications in education by (Chen et al., 2020) identifies natural language processing (NLP), intelligent tutoring systems, and learning analytics as key areas where artificial intelligence is transforming educational practices. This contextual foundation supports the increasing use of sentiment classification in understanding learner perceptions. For example, recent studies have analyzed public discourse on education through sentiment analysis applied to social media platforms such as Twitter, providing insights into broader societal perspectives on education systems (Mouronte-López et al., 2023).

Prior studies have applied sentiment classification to student feedback in order to uncover student perceptions related to courses, instructors, and learning environments, particularly within higher education contexts (Brânzilă, 2024; Dake & Gyimah, 2023; Grimalt-Álvaro & Usart, 2024; Shaik et al., 2023). In many of these studies, classical machine learning models are preferred due to their interpretability, relatively low computational cost, and suitability for institutional settings where resources may be constrained. While these approaches have demonstrated practical value for large-scale feedback analysis, they are often evaluated primarily in terms of predictive performance, with more limited discussion of how such models support pedagogical interpretation or institutional decision-making processes.

Techniques such as Logistic Regression and Support Vector Machines have been effectively used in educational settings to predict sentiment polarity in course evaluations and discussion forums (Koufakou, 2024; Toçoğlu & Onan, 2021). More recently, ensemble methods like Random Forests and XGBoost have been employed for their ability to capture non-linear relationships and interactions among features (Fatima et al., 2023; Kumar et al., 2024). Recent studies have further advanced sentiment analysis in educational contexts by incorporating sophisticated neural architectures. For instance, a hybrid BERT–LSTM–CNN model was proposed to analyze student reviews in MOOCs, achieving improved performance in capturing emotional nuances in online learning environments (Baqach & Battou, 2024). Similarly, the integration of knowledge graphs into sentiment analysis workflows has enhanced e-learning evaluation by bridging semantic representation with AI-driven classification techniques (Yi et al., 2025). In parallel, deep learning architectures, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, have demonstrated superior performance in capturing contextual nuances in student-written text, especially when dealing with long or unstructured feedback (Iparraguirre-Villanueva et al., 2023; Reddy et al., 2022; Shen et al., 2020). Preprocessing and feature representation are essential for the performance of sentiment classifiers. While TF-IDF has long been a standard for traditional models due to its simplicity and effectiveness (Abubakar et al., 2022), word embeddings such as Word2Vec and GloVe have shown advantages in preserving semantic relationships, particularly for neural models (Alamoudi & Alghamdi, 2021; Onan, 2021). Hybrid approaches that combine traditional and deep learning models are also gaining traction in the educational domain (Mary & Claral, 2025; Viveka & Priya, 2024). Furthermore, addressing class imbalance is critical, especially in binary sentiment tasks where negative feedback is often underrepresented. Recent studies have also applied transformer-based architectures such as BERT and RoBERTa for educational sentiment analysis tasks, demonstrating superior performance due to their ability to capture contextual semantics, although their high computational requirements can limit deployment in resource-constrained settings (Devlin et al., 2019; Y. Liu et al., 2019). Synthetic sampling methods such as SMOTE have been proposed to improve recall and fairness across classes, although their effectiveness varies by model architecture and dataset characteristics (Chawla et al., 2002; Sakho et al., 2024).

Beyond algorithmic design, studies have emphasized the importance of optimizing classifier thresholds to better align predictions with institutional needs, particularly when prioritizing early detection of student dissatisfaction (Hancock et al., 2022; Rodríguez Velasco et al., 2023). While default thresholds (e.g., 0.5) are commonly used, threshold calibration has proven effective in educational sentiment contexts to balance precision and recall. Collectively, this body of work demonstrates significant methodological progress in the application of sentiment analysis to educational data, yet it remains largely oriented toward performance-driven evaluation rather than institutional applicability or pedagogical decision-making.

Despite the growing interest in educational sentiment analysis, relatively few studies have examined authentic Spanish-language student feedback within Latin American higher education institutions. Existing work in this region is often limited in scale or scope and rarely addresses the practical constraints—such as limited computational resources or interdisciplinary expertise—that shape real-world institutional adoption.

Overall, existing research demonstrates the growing relevance of sentiment analysis as a tool for extracting insights from student feedback in educational settings. Prior studies have explored a range of machine learning and deep learning approaches, often reporting promising predictive performance. However, much of this literature remains focused on algorithmic optimization or isolated case studies, with limited attention to how methodological choices support institutional decision-making processes in higher education. Moreover, there is a noticeable lack of systematic evaluations conducted using authentic Spanish-language student feedback from underrepresented regions such as Latin America, particularly under conditions of limited computational and human resources. These gaps highlight the need for frameworks that move beyond performance comparison to emphasize interpretability, scalability, and practical applicability—motivating the context-aware, multi-phase approach proposed in this study.

3. Context and Dataset

3.1. Institutional Context

The dataset originates from a private university in Peru with multiple campuses, where student satisfaction surveys are conducted to assess academic quality, faculty performance, curriculum effectiveness, and student support services. Unlike traditional evaluations that primarily focus on administrative aspects, this dataset provides insights into pedagogical factors that directly impact student engagement and learning outcomes.

Sentiment analysis of this feedback enables institutions to identify key drivers of student motivation, instructional effectiveness, and areas that need improvement, ultimately supporting the development of targeted educational strategies. However, several challenges arise in the Peruvian context, including linguistic diversity (Spanish and indigenous languages), resource limitations (restricted access to AI infrastructure), and cultural nuances in sentiment expression, all of which can influence the precision of sentiment classification models.

Additionally, recent literature has emphasized how the COVID-19 pandemic transformed student feedback dynamics, further underscoring the importance of timely, automated analysis tools in hybrid or online learning environments (Salas-Pilco et al., 2022).

https://doi.org/10.3390/educsci16030399

Addressing these challenges is crucial for designing an effective sentiment analysis framework that translates student feedback into actionable insights. Using data-driven decision making, universities can improve teaching methodologies, optimize curriculum structures, and implement support mechanisms that foster an engaging and inclusive learning environment.

3.2. Dataset and Data Collection

This study is based on a dataset comprising 23,168 open-ended comments provided by undergraduate students (average age between 18 and 22) across four university campuses in Peru. The data were collected during two academic semesters in 2019 as part of a student satisfaction survey designed to capture students' perceptions not only of university services, but also of their academic experiences and engagement levels.

Although the data were collected in 2019, the structure of the feedback instruments and the institutional evaluation context remain representative of current higher education practices in Latin American universities.

A web-based application was developed specifically for data collection and integrated into the virtual classrooms of all faculties. The application remained active for two weeks in the middle of each semester, allowing broad participation from students in different academic programs. This approach ensured a representative collection of student feedback regarding various aspects of university life.

Each comment was manually annotated by faculty members assigned to their respective academic departments using a binary labeling scheme: Positive (comments that express satisfaction, praise, or approval toward a course, faculty member, or learning experience) and Negative (comments that express dissatisfaction, criticism, or concerns regarding any academic aspect). Neutral or ambiguous comments were discarded during preprocessing.

This decision was made to reduce annotation ambiguity and to focus the analysis on sentiment signals that are more directly actionable for institutional and pedagogical decision-making.

In addition to sentiment polarity, annotations included thematic categorization based on message content. The institutional categories used for this purpose were General Perception, Teachers, Academic Strategy, and Academic Experience, encompassing a broad range of academic and support-related dimensions.

To illustrate the nature of the annotations, Table 1 presents a sample of real student comments along with their sentiment labels in Spanish and English, their translations, and the institutional categories assigned. These examples help clarify how faculty members interpreted and categorized the feedback across diverse academic themes.

For model development and evaluation, the dataset was first divided into two primary subsets following the experimental design shown in Figure 1: an 85% experimental set and a 15% held-out final test set. The experimental set (85%) was subsequently subdivided into training (70% of the full dataset; 16,217 comments) and validation (15% of the full dataset; 3476 comments) subsets. The training subset was used for model learning, the validation subset for hyperparameter tuning and model selection, and the final 15% test set was reserved exclusively for unbiased performance evaluation.

Table 1. Examples of Student Comments Annotated by Faculty Members.

Comment (Spanish)	Label (Spanish)	Comment (English)	Label (English)	Institutional Category (English)
Los docentes explican bien y son muy pacientes al responder dudas.	Positivo	Teachers explain well and are very patient when answering questions.	Positive	Teachers
La estrategia del curso me permitió reforzar conocimientos clave.	Positivo	The course strategy helped me reinforce key knowledge.	Positive	Academic Strategy
Algunos docentes no están preparados y no manejan bien el tema.	Negativo	Some teachers are not prepared and do not manage the topic well.	Negative	Teachers
No hay coordinación entre los contenidos y lo que se evalúa.	Negativo	There is no coordination between the content and what is assessed.	Negative	Academic Strategy

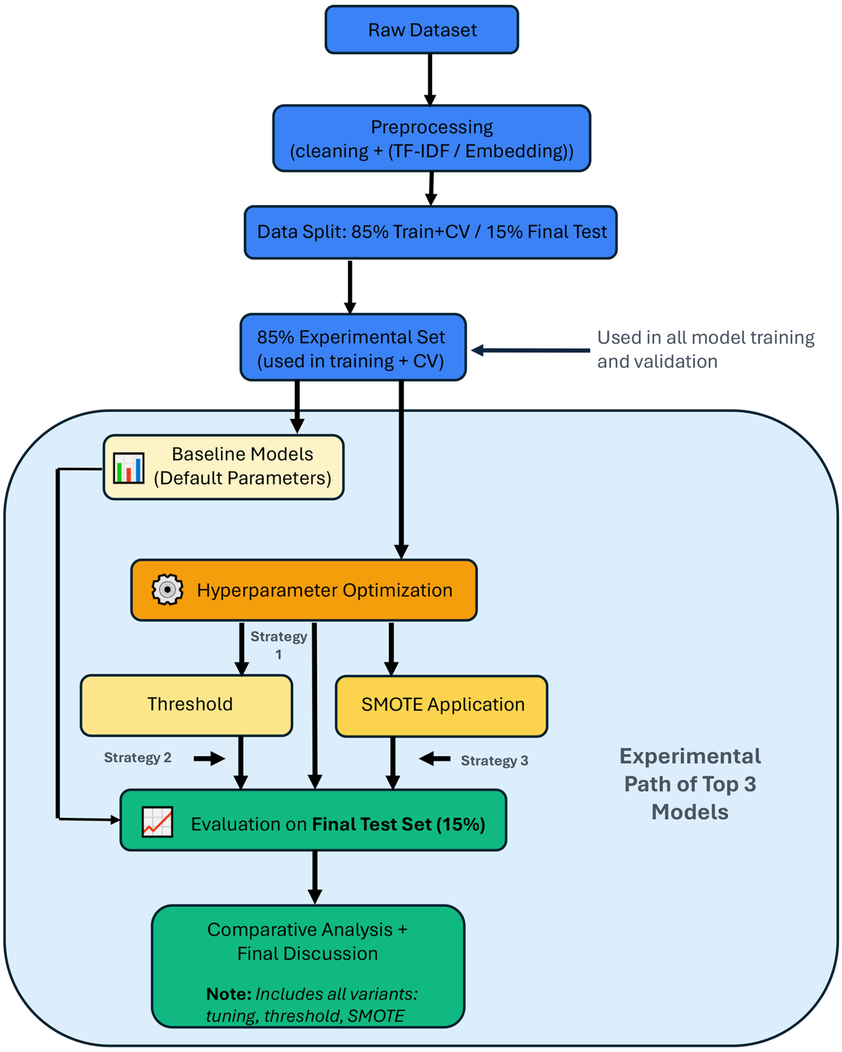

Figure 1. Experimental design for sentiment classification evaluation and optimization. Blue boxes represent preprocessing stages, green boxes denote model training and selection, and orange boxes indicate optimization strategies.

This hierarchical splitting strategy—resulting in an effective 70/15/15 partition—differs slightly from the commonly used 70/20/10 scheme but is widely accepted in applied machine learning, particularly when working with imbalanced classes and calibration procedures. It ensures sufficient data for optimization while preserving a strictly unseen test set for final evaluation (Géron, 2022; Goodfellow et al., 2016).

To facilitate a deeper analysis and highlight key educational insights, the student comments were grouped into refined thematic subcategories derived from contextual interpretation and linguistic patterns. While the original manual annotation followed broader institutional dimensions (General Perception, Teachers, Academic Strategy, and Academic Experience), these were further refined for analytical purposes into pedagogically oriented themes. The categories presented in Table 2 therefore represent an operational refinement of the initial institutional labels rather than a replacement of them. This hierarchical categorization enabled a more nuanced exploration of the relationship between sentiment expression and specific educational dimensions of learning quality.

Table 2. Representative student comments grouped by educational themes derived from linguistic patterns and institutional categories.

Student Comment	Category
Actually, everything. The professors, service, poor management.	Teaching Methodology
The teaching quality is very good.	Academic Quality
Academic rigor and level.	Academic Model
Development of syllabi consistent with the faculty.	Curriculum and Course Planning
The library should be bigger to provide more comfort.	Library and Study Resources

By analyzing these patterns of sentiment, the study aims to bridge the gap between student feedback and actionable improvements in educational quality. Understanding how students articulate their experiences provides institutions with a data-driven foundation to enhance teaching practices, improve academic design, and implement targeted support interventions.

All data were anonymized prior to analysis, and no personally identifiable information (PII) was collected. The dataset, derived from institutional feedback surveys, was processed in accordance with ethical standards for educational data use.

The study involves minimal risk to participants, as it relies exclusively on anonymized, retrospective educational data collected for institutional quality assurance purposes.

4. Methodology

The methodology of this study was designed to evaluate and optimize the performance of sentiment analysis models applied to student feedback, with the goal of informing data-driven educational strategies in Latin America. The framework followed a structured, multi-phase design consisting of three main stages: data preparation, model training and enhancement, and performance evaluation.

The methodological design prioritizes transparency and reproducibility, ensuring that the framework can be understood and implemented by interdisciplinary institutional teams rather than machine learning specialists alone.

Figure 1 illustrates the overall experimental design.

4.1. Data Preparation

To prepare the dataset for sentiment classification, several preprocessing steps were applied to clean and normalize the text. These included converting all text to lowercase, removing punctuation and special characters, eliminating stopwords, and applying lemmatization to reduce words to their base forms. These operations ensured consistency and reduced noise.

For the traditional models (Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, and XGBoost), feature extraction was conducted using Term Frequency–Inverse Document Frequency (TF-IDF) (Abubakar et al., 2022). This approach provided sparse vector representations where term weights reflected their importance relative to the corpus. For the LSTM model, tokenized and padded sequences were passed through an embedding layer to capture semantic and syntactic relationships (Alamoudi & Alghamdi, 2021; Onan, 2021).

This differentiated preprocessing strategy reflects common institutional scenarios in which traditional models are favored for rapid deployment, while deep learning models are reserved for more detailed exploratory analysis.

The dataset exhibited a mild class skew, with 57.45% positive and 42.55% negative instances. No balancing was applied at this stage in order to preserve the natural distribution of student feedback, reflecting realistic institutional conditions where sentiment data are rarely perfectly balanced. These preparations ensured that each model received input aligned with its design characteristics.

4.2. Model Training and Enhancement

Six classification models were selected based on their relevance in natural language processing and educational sentiment analysis (Brânzilă, 2024; Toçoğlu & Onan, 2021): Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, XGBoost, and LSTM. Each model was first trained using default hyperparameters, with 85% of the data used for training and internal validation, and 15% reserved for final evaluation.

Based on baseline performance, three models—Logistic Regression, XGBoost, and LSTM—were selected for further enhancement. Each enhancement strategy was selected to reflect a realistic institutional decision point rather than to maximize performance at all costs.

Optimization was carried out through three distinct strategies, as illustrated in Figure 1, which visually outlines the experimental pipeline for clarity. First, hyperparameter tuning was performed using Bayesian Optimization combined with five-fold cross-validation. The search space included model-specific parameters such as regularization strength and learning rate (Garnett, 2023). Second, threshold adjustment was applied to calibrate decision boundaries, improving the detection of negative sentiment by maximizing the F1-score—particularly relevant in the context of quality assurance in education (Hancock et al., 2022; Rodríguez Velasco et al., 2023). Threshold calibration is particularly relevant in educational settings, where the early identification of negative sentiment may be more valuable than marginal gains in overall accuracy. Third, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002; Sakho et al., 2024) was used to address class imbalance by generating synthetic samples for the minority class. These three enhancement strategies were implemented independently to isolate and evaluate their individual effects on model robustness and generalization.

4.3. Evaluation Strategy

Model evaluation was carried out using the 15% held-out test set, untouched during training and optimization phases. Metrics included accuracy, precision, recall, and F1-score, all computed as weighted averages to address class imbalance. The evaluation protocol was consistent across all model variants, allowing fair comparisons. Five-fold cross-validation was used during optimization to ensure reliability. Final model performance was assessed on the unseen test set to reflect real-world deployment conditions. These metrics were chosen for their interpretability and alignment with institutional priorities focused on identifying dissatisfaction and supporting continuous improvement in higher education (Dake & Gyimah, 2023; Grimalt-Álvaro & Usart, 2024). This evaluation design mirrors real-world deployment conditions, where models are trained once and subsequently applied to unseen student feedback streams.

During manuscript preparation, the authors used an AI-assisted writing tool to help improve the clarity and flow of the text. All content was carefully reviewed, revised, and approved by the authors to ensure its accuracy and originality.

5. Results and Discussion

This section presents and interprets the findings of the model evaluation, with emphasis on comparative insights and practical implications. To avoid redundancy, details of model configurations and methodology have been streamlined. Only the most relevant tables and figures are retained.

5.1. Performance Overview and Model Selection

This subsection provides an initial comparative overview of baseline model performance, serving as a filtering stage to identify classifiers that balance predictive quality with institutional feasibility. Rather than prioritizing maximum accuracy alone, this stage emphasizes models that offer stable performance while remaining interpretable and computationally efficient for deployment in higher education quality assurance contexts.

Six classification models were evaluated under default hyperparameters using five-fold cross-validation. Logistic Regression, Naïve Bayes, and LSTM achieved the highest F1-scores, indicating reliable performance across sentiment classes. Random Forest and XGBoost followed closely, while SVM offered competitive but less distinctive results.

Despite its strong metrics, Naïve Bayes relies on simplifying assumptions that limit its generalizability to nuanced student feedback (Jurafsky & Martin, 2026). SVM showed no substantial gains in precision or recall and is computationally more intensive (Cipolla & Gondzio, 2022). Consequently, Logistic Regression, XGBoost, and LSTM were selected for further enhancement, given their balanced trade-offs between performance, efficiency, and interpretability (Dake & Gyimah, 2023; Wahyuddin et al., 2025).

Overall, these baseline results indicate that strong predictive performance must be considered alongside practical constraints. Models exhibiting marginal performance advantages but requiring substantially higher computational or technical resources may offer limited added value for routine institutional use.

5.2. Enhancement Results: Optimization, Thresholding, and SMOTE

This subsection examines how commonly used enhancement strategies—Bayesian optimization, threshold calibration, and class balancing—affect model behavior beyond raw performance gains. The analysis focuses on understanding how these strategies influence model reliability and applicability in realistic institutional deployment scenarios.

Bayesian optimization led to performance improvements in all three selected models. LSTM yielded the highest F1-score, confirming its suitability for capturing sequential sentiment cues. Logistic Regression and XGBoost showed stable gains. Threshold calibration provided slight recall boosts for LSTM, which may aid in identifying negative sentiment—a key concern for educational quality monitoring (Hancock et al., 2022).

In contrast, SMOTE-based balancing introduced marginal performance declines for Logistic Regression and XGBoost, and only modest improvement for LSTM. These results align with prior findings on the limited utility of synthetic oversampling in moderately imbalanced datasets (Sakho et al., 2024). Thus, SMOTE is not universally beneficial and should be applied with caution.

While oversampling techniques such as SMOTE improved recall in several cases, these gains were not uniform across all models. This suggests that class balancing should be applied selectively, particularly in educational datasets where class imbalance is moderate and interpretability remains a priority. Taken together, these findings highlight that enhancement strategies should be evaluated not only for their ability to improve metrics, but also for their impact on model stability, interpretability, and alignment with institutional objectives.

5.3. Final Test Evaluation and Comparative Insights

Final evaluation was conducted on the 15% held-out test set. Among all configurations, LSTM optimized with SMOTE achieved the best overall F1-score (0.7589), followed closely by optimized Logistic Regression (0.7587) and XGBoost (0.7417). Default Logistic Regression also remained competitive, reaffirming its value as a lightweight, interpretable model.

To further illustrate the comparative effects of the enhancement strategies, Figure 2 presents a bar chart of F1-scores across baseline and optimized configurations for the three best-performing models. This visual complement to Table 3 helps to more clearly distinguish performance trends across strategies. Notably, LSTM with SMOTE yielded the highest F1-score, while hyperparameter optimization (HO) also showed consistent improvements across models. The chart reinforces that no single strategy dominates across classifiers and highlights how model-strategy alignment is context dependent.

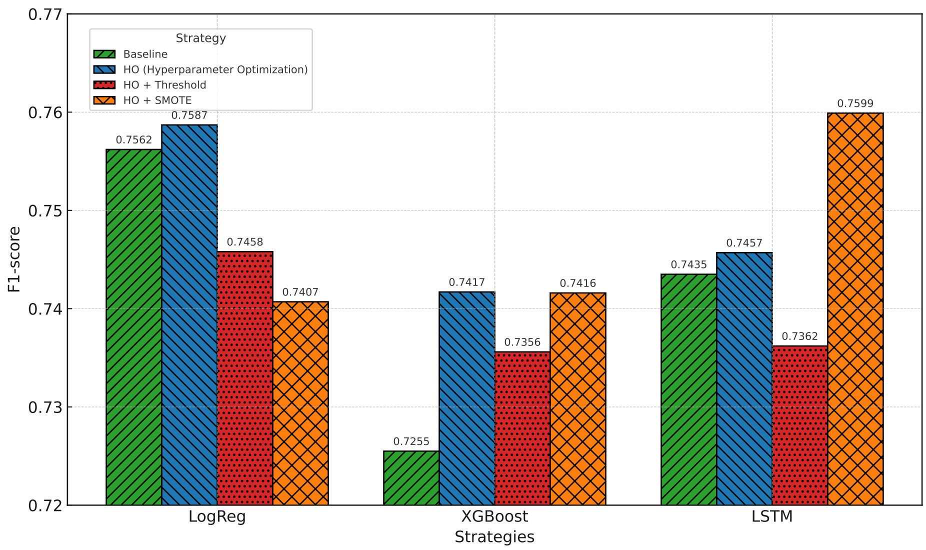

Figure 2. F1-Score comparison across enhancement strategies.

To further examine model behavior on unseen data, we present the confusion matrix for the best-performing configuration of each classifier. These visualizations offer insights into how each model distinguishes between positive and negative classes, highlighting strengths and common misclassification patterns that may affect practical deployment in educational settings.

Figure 3 shows the results for Logistic Regression using optimized hyperparameters. The model demonstrates balanced performance, with a slight tendency to misclassify negative sentiments, a common limitation in models trained on imbalanced datasets.

Figure 3. Confusion matrix for Logistic Regression (Optimized) on the 15% test set.

Figure 4 presents the matrix for the optimized XGBoost model. While it maintains strong performance for the positive class, its precision on the negative class is slightly lower than that of Logistic Regression. Nevertheless, the model offers a balanced trade-off between complexity and accuracy.

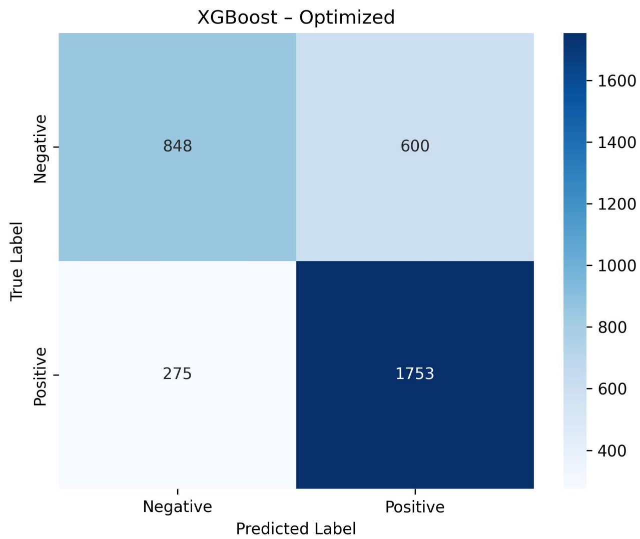

Figure 4. Confusion matrix for XGBoost (Optimized) on the 15% test set.

Finally, Figure 5 displays the matrix for the LSTM model trained with optimized architecture and SMOTE-based class balancing. This model shows the best performance in both sentiment classes, with low misclassification rates. Its sequential structure appears especially effective in capturing sentiment nuances within student comments.

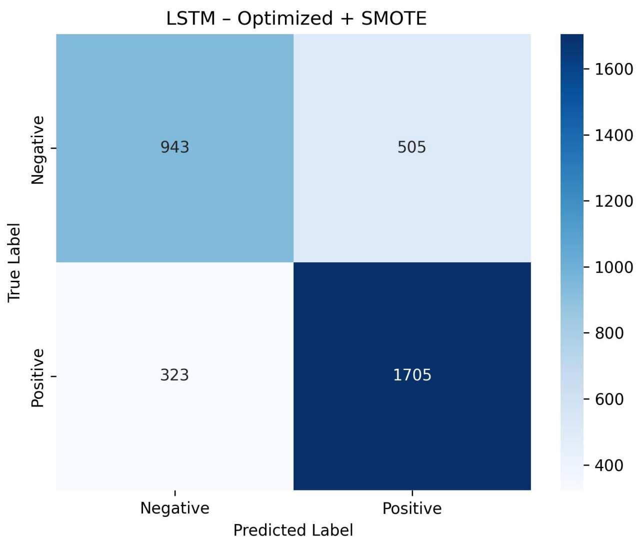

Figure 5. Confusion matrix for LSTM (Optimized + SMOTE) on the 15% test set.

To complement this class-level analysis, Table 3 summarizes the final evaluation metrics—accuracy and F1-score—across all model configurations. Although differences are numerically small, they underscore the need to consider trade-offs such as interpretability and computational cost during model selection. As a visual aid, Figure 2 presents the comparative trends in F1-score for baseline and enhanced setups.

These results provide a solid foundation for reflecting on model applicability in real-world institutional settings, as discussed in the following section.

Table 3. Final Test Metrics (15%) for All Models and Configurations. **Bolded values** indicate best performance per metric.

Model + Config	Accuracy	Precision	Recall	F1-Score
LogReg − Default Params	0.7604	0.7598	0.7604	0.7562
LogReg − Optimized	0.7595	0.7584	0.7595	0.7587
XGBoost − Optimized	0.7483	0.7492	0.7483	0.7417
LSTM − Optimized + SMOTE	**0.7618**	**0.7604**	**0.7618**	**0.7589**

From an institutional perspective, the relatively small performance differences observed across optimized models reinforce the importance of selection criteria beyond peak accuracy. Factors such as transparency, maintainability, and ease of integration into existing quality assurance workflows become decisive when deploying sentiment analysis systems at scale.

These results underscore that, in applied educational contexts, model selection should be guided by contextual suitability rather than marginal performance improvements alone.

From a practical instructional perspective, the comparison of multiple models in this study is not intended to emphasize model superiority, but rather to ensure the robustness and reliability of the proposed framework. While performance metrics provide methodological validation, the primary educational value lies in the interpretability of aggregated sentiment patterns that can inform instructional and institutional decision-making. In educational settings, such interpretable trends are often more actionable than marginal improvements in accuracy, as they support the identification of areas requiring pedagogical attention. A more detailed qualitative examination of concrete student feedback excerpts is therefore framed as a natural direction for future research.

https://doi.org/10.3390/educsci16030399

5.4. Implications for Educational Deployment

While the models evaluated in this study are widely established in natural language processing, their comparative performance has rarely been assessed on authentic Spanish-language student feedback within Latin American universities. Accordingly, the primary contribution of this work is not the introduction of new algorithms, but the validation of a replicable, context-aware framework that supports informed educational decision-making under realistic institutional constraints.

Logistic Regression offers strong interpretability and efficiency, making it ideal for real-time applications in institutional dashboards. XGBoost balances performance with moderate complexity but may not justify the added tuning in all contexts. LSTM, while resource-intensive, excels when deeper analysis of narrative feedback is needed. Institutions can adopt a tiered approach: lightweight models for large-scale monitoring, and deep models for targeted, high-sensitivity applications.

Overall, the findings suggest that no single enhancement strategy consistently outperforms the others across all models. While optimization and thresholding contribute measurable improvements, class balancing with SMOTE produces mixed results depending on the classifier architecture. While SMOTE can be beneficial in scenarios where minority-class recall is prioritized, its effectiveness is highly context-dependent and must be evaluated in relation to model architecture and institutional deployment constraints. These outcomes reinforce the importance of aligning model selection and enhancement strategies with institutional objectives, data characteristics, and deployment constraints.

5.4.1. Integration of Sentiment Analysis into Institutional Quality Processes

To move from model evaluation to institutional impact, sentiment analysis must be embedded within existing quality assurance and feedback management processes rather than treated as an isolated analytical task.

To translate technical findings into actionable educational improvements, it is essential to frame sentiment analysis not just as a data processing task but as part of a broader institutional ecosystem. The proposed framework offers a pathway for integrating automated feedback analysis into continuous quality assurance (QA) workflows, helping universities become more responsive, student-centered, and evidence-informed.

Figure 6 presents a conceptual model for institutional deployment. This structure outlines how sentiment classification outputs can support key processes in academic management and quality enhancement:

- Real-time Feedback Monitoring: Automated dashboards can alert faculty and administrators to recurring issues or shifts in student sentiment during the term.
- Early Warning and Intervention: Negative sentiment patterns—when promptly detected—can trigger academic or psychological support interventions.
- Data-informed Pedagogical Adjustments: Aggregated insights can guide improvements in course content, instructional design, and assessment practices.
- Evidence for Accreditation and Reporting: Sentiment trends can complement traditional indicators (e.g., grades, dropout rates) as part of quality assurance documentation and reporting to accrediting bodies.
- Institutional Strategy Alignment: Feedback-driven insights can support decision-making at curricular, departmental, and institutional levels, aligning actions with strategic educational goals.

https://doi.org/10.3390/educsci16030399

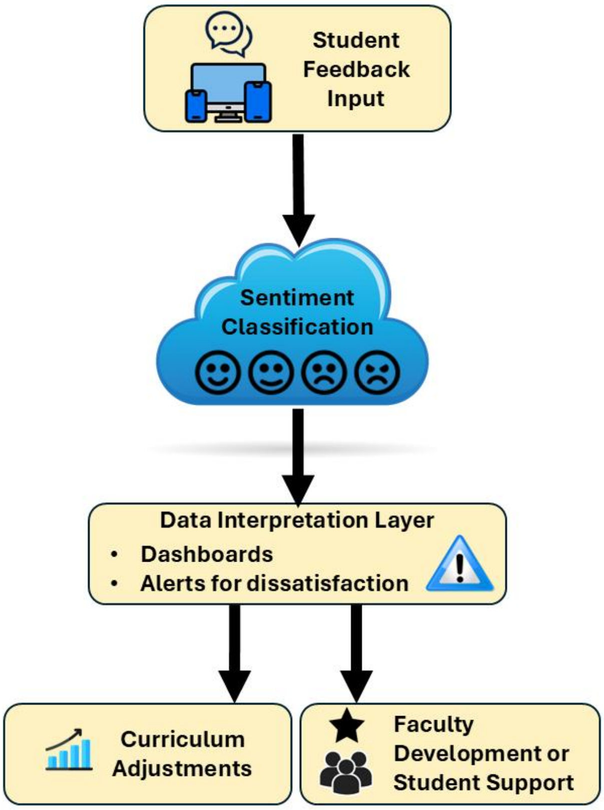

Figure 6. Integration of sentiment analysis into institutional feedback and decision-making processes.

This integration does not require full automation of decisions but rather enhances existing institutional workflows with AI-generated signals. The architecture illustrated in Figure 6 supports scalable and ethical application of sentiment analysis aligned with regional accreditation frameworks and student engagement priorities in Latin American higher education.

In addition to the quantitative evaluation based on accuracy and F1-score, we incorporated two qualitative dimensions—interpretability and computational simplicity—to enable a more comprehensive comparison of the models under realistic deployment conditions. These dimensions were scored on a normalized 1–5 scale, where higher values represent more interpretable models and lower computational demands. Interpretability was rated based on how easily a model's predictions can be understood and explained by non-technical users, following criteria outlined in (Swamy et al., 2023). Simplicity, on the other hand, reflects the relative ease of training and deploying each model, particularly in resource-constrained academic environments, consistent with (Q. Liu et al., 2024).

Among the models evaluated, Logistic Regression achieved the highest marks in both interpretability and simplicity, owing to its transparent structure and minimal computational requirements. XGBoost offered a balance between performance and usability, with moderate interpretability through feature importance analysis. In contrast, LSTM received lower scores on these dimensions due to its complexity, opaque decision process, and reliance on deep learning infrastructure.

To visually synthesize these trade-offs, Figure 7 presents a radar chart that integrates both the quantitative metrics (accuracy and F1-score) and qualitative assessments. This visualization highlights how model selection in educational applications may depend not only on raw performance but also on practical considerations such as interpretability and ease of integration into institutional workflows.

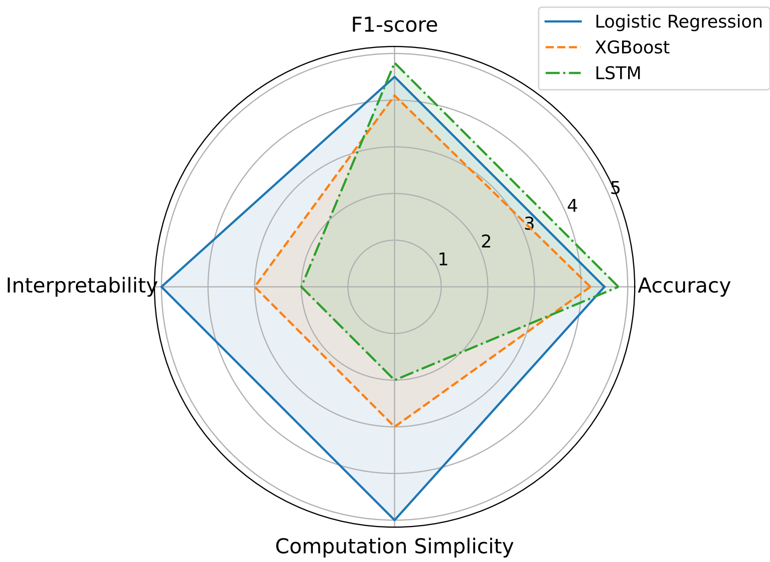

Figure 7. Radar chart comparing Logistic Regression, XGBoost, and LSTM across four key dimensions: Accuracy, F1-score, Interpretability, and Computation Cost (Inverted).

This integrated approach reinforces the practical utility of sentiment analysis in quality assurance ecosystems, positioning machine learning not as a replacement for human judgment, but as a scalable complement that empowers institutions to act on student voice in timely and meaningful ways.

6. Conclusions and Future Work

This study presented a context-aware framework for sentiment analysis of student feedback designed to support institutional decision-making in higher education. Rather than introducing novel algorithms, the work systematically evaluated how established machine learning and deep learning models behave under realistic constraints commonly faced by Latin American universities, including limited computational resources and the need for interpretability. The results demonstrate that marginal differences in predictive performance across optimized models often translate into negligible practical advantages at the institutional level. In this context, simpler and more transparent models can offer comparable value to more complex architectures, particularly when ease of deployment, maintainability, and alignment with quality assurance processes are prioritized. Among the evaluated models, LSTM achieved the highest overall performance with an accuracy of 76.18% and an F1-score of 75.89%, particularly when combined with class balancing techniques. Logistic Regression, while simpler, also demonstrated competitive performance (accuracy: 75.95%, F1-score: 75.87%) and stood out for its interpretability and operational efficiency. XGBoost offered modest improvements (accuracy: 74.83%, F1-score: 74.17%) after tuning, though it required greater computational complexity.

Not all enhancement strategies yielded uniform benefits. Hyperparameter optimization consistently improved performance, while threshold calibration and SMOTE produced mixed results depending on the model and dataset characteristics. These findings highlight the importance of aligning model selection and enhancement strategies with real-world deployment conditions, institutional constraints, and educational goals.

For universities seeking scalable, transparent, and actionable tools, interpretable models such as Logistic Regression may offer the best trade-off between usability and predictive power to support continuous quality improvement and data-informed pedagogical adjustments. Conversely, more complex architectures like LSTM can support advanced tasks such as early identification of dissatisfaction patterns or longitudinal tracking of sentiment shifts, contributing to more responsive and student-centered educational strategies.

In this sense, the proposed framework contributes to strengthening institutional mechanisms for educational quality assurance and continuous improvement, offering practical evidence to guide policy and curriculum development while aligning with the broader objectives of educational innovation and accountability.

While this study focuses on aggregated sentiment patterns to support institutional and instructional decision-making, a more fine-grained qualitative analysis of concrete student feedback excerpts falls beyond its current scope and is therefore identified as a relevant direction for future research.

One limitation of this study is that the dataset used was collected in 2019. However, its linguistic patterns, thematic categories, and institutional context remain highly representative of current Latin American higher education settings, where curricular structures and modes of feedback collection have changed gradually. Future work will incorporate more recent post-pandemic datasets to explore shifts in student sentiment expression under hybrid and online learning modalities.

By framing sentiment analysis as an operational support tool rather than a purely technical classification task, this work contributes to bridging the gap between learning analytics research and actionable educational practice. The proposed framework provides institutions with a replicable approach to extracting meaningful insights from student feedback to inform pedagogical strategies and continuous improvement initiatives.

In addition, the framework preserves a clear hierarchical structure between the original institutional annotation dimensions and the refined pedagogical subcategories used for thematic interpretation. This distinction ensures methodological transparency and supports the translation of sentiment outputs into institutionally meaningful quality indicators without altering the foundational labeling scheme.

Future work will extend the proposed framework toward multi-class and aspect-based sentiment analysis, enabling institutions to generate more granular and targeted insights to support teaching and learning enhancement. While transformer-based architectures such as BERT and RoBERTa have demonstrated strong performance in natural language processing tasks, this study deliberately prioritized models with lower computational requirements to ensure feasibility in institutions with limited AI infrastructure.

Accordingly, future research may explore the integration of multilingual transformers and contextual embeddings in a controlled manner, with particular attention to deployment cost, interpretability trade-offs, and alignment with institutional capacities. In parallel, the incorporation of explainable AI (XAI) techniques could further enhance transparency and trust in model outputs, supporting their use by academic leaders and decision-makers (Q. Liu et al., 2024; Swamy et al., 2023). Cross-institutional validation across Latin America would also contribute to assessing generalizability and identifying region-specific patterns relevant to inclusive educational improvement.

Overall, while this work does not aim to advance new machine learning architectures, it contributes a context-aware and resource-sensitive framework that combines established techniques for educational quality improvement. This pragmatic focus strengthens the potential for adoption by higher education institutions in Latin America seeking scalable and interpretable AI-driven tools for strategic decision-making.

Author Contributions: Conceptualization, A.P.-B. and J.O.C.; methodology and validation, A.P.-B. and R.B.; software and experimental scripts, A.P.-B. and L.C.S.; formal analysis, A.P.-B. and W.S.; investigation, A.P.-B., L.C.S. and W.S.; data curation, J.O.C., L.C.S. and W.S.; writing—original draft preparation, A.P.-B. and R.B.; writing—review and editing, R.B., L.C.S. and W.S.; visualization, A.P.-B.;

supervision and project administration, A.P.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted with institutional support from the Tecnológico Nacional de México/Instituto Tecnológico de Matamoros and funded by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SeCiHTI), México.

Institutional Review Board Statement: The study was conducted in accordance with the ethical guidelines of the participating Peruvian University, where the data were collected as part of official institutional student satisfaction surveys. The research protocol was reviewed and approved by the university's Academic Quality and Research Committee. All data were anonymized prior to analysis to ensure compliance with ethical standards for educational research.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. Participation in the institutional satisfaction survey was voluntary, and responses were collected anonymously through the university's academic platform. No personally identifiable information was gathered.

Data Availability Statement: The dataset used in this study is publicly available in Zenodo at https://doi.org/10.5281/zenodo.18797217.

Acknowledgments: The authors would like to thank the faculty members involved in the manual annotation process for their valuable contribution to dataset preparation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, *4*(1), 27–33. [CrossRef]

Alamoudi, E. S., & Alghamdi, N. S. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, *30*(2–3), 259–281. [CrossRef]

Baqach, A., & Battou, A. (2024). A new sentiment analysis model to classify students' reviews on MOOCs. *Education and Information Technologies*, *29*(13), 16813–16840. [CrossRef]

Barbosa, M. L. d. O., & Marín-Suelves, D. (2024). Content and sentiment analysis of autobiographical narratives of experienced and well-evaluated teachers in Spain. *Education Sciences*, *14*(6), 642. [CrossRef]

Brânzilă, C. (2024). Sentiment analysis as a tool for optimizing educational strategies and business communication. *Linguaculture*, *15*(Special Issue), 83–98. [CrossRef]

Cechinel, C., Ochoa, X., Lemos dos Santos, H., Carvalho Nunes, J. B., Rodés, V., & Marques Queiroga, E. (2020). Mapping learning analytics initiatives in Latin America. *British Journal of Educational Technology*, *51*(4), 892–914. [CrossRef]

Chawla, N., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. [CrossRef]

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264–75278. [CrossRef]

Cipolla, S., & Gondzio, J. (2022). Training very large scale nonlinear SVMs using Alternating Direction Method of Multipliers coupled with the Hierarchically Semi-Separable kernel approximations. *EURO Journal on Computational Optimization*, *10*, 100046. [CrossRef]

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies*, *28*(4), 4629–4647. [CrossRef]

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Fatima, S., Hussain, A., Amir, S. B., Ahmed, S. H., Aslam, S. M. H. (2023). XGBoost and random forest algorithms: An in depth analysis. *Pakistan Journal of Scientific Research*, *3*(1), 26–31. [CrossRef]

Garnett, R. (2023). *Bayesian optimization* (1st ed.). Cambridge University Press.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). MIT Press. Available online: https://www.deeplearningbook.org/ (accessed on 25 March 2025).

Grimalt-Álvaro, C., & Usart, M. (2024). Sentiment analysis for formative assessment in higher education: A systematic literature review. *Journal of computing in higher education*, *36*(3), 647–682. [CrossRef]

Hancock, J., Johnson, J. M., & Khoshgoftaar, T. M. (2022). A comparative approach to threshold optimization for classifying imbalanced data. In *2022 IEEE 8th international conference on collaboration and internet computing (CIC)* (pp. 135–142). IEEE.

Iparraguirre-Villanueva, O., Guevara-Ponce, V., Ruiz-Alvarado, D., Beltozar-Clemente, S., Sierra-Liñan, F., Zapata-Paulini, J., & Cabanillas-Carbonell, M. (2023). Text prediction recurrent neural networks using long shortterm memory-dropout. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*, 1758–1768. [CrossRef]

Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, with language models* (3rd ed.). Stanford University. Available online: https://web.stanford.edu/~jurafsky/slp3/ (accessed on 7 January 2026).

Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. *Education and Information Technologies*, *29*(3), 2973–2997. [CrossRef]

Kumar, M., Singh, N., Wadhwa, J., Singh, P., Kumar, G., & Qtaishat, A. (2024). Utilizing random forest and XGBoost data mining algorithms for anticipating students' academic performance. *International Journal of Modern Education and Computer Science*, *16*(2), 29–44. [CrossRef]

Liu, Q., Pinto, J. D., & Paquette, L. (2024). Applications of explainable AI (XAI) in education. In *Trust and inclusion in AI-mediated education: Where human learning meets learning machines* (pp. 93–109). Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv*, arXiv:1907.11692.

Mary, T., & Claral, A. (2025). Hybrid deep learning model to predict students' sentiments in higher educational institutions. *International Journal of Interactive Mobile Technologies*, *19*(1), 46–61.

Masiello, I., Mohseni, Z., Palma, F., Nordmark, S., Augustsson, H., & Rundquist, R. (2024). A current overview of the use of learning analytics dashboards. *Education Sciences*, *14*(1), 82. [CrossRef]

Mouronte-López, M. L., Ceres, J. S., & Columbrans, A. M. (2023). Analysing the sentiments about the education system trough Twitter. *Education and Information Technologies*, *28*(9), 10965–10994. [CrossRef] [PubMed]

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), e5909. [CrossRef]

Patil, P., Kolamkar, N., Mainkar, M., & Hande, Y. (2024). NLP-based system real-time analysis of students textual feedback to enhance online class session insight. In *2024 IEEE international conference on blockchain and distributed systems security (ICBDS)* (pp. 1–7). IEEE. [CrossRef]

Pereiro, E., Montaldo, M., Koleszar, V., & Urruticoechea, A. (2022). *Computational thinking, artificial intelligence and education in Latin America*. UNESCO. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000381761 (accessed on 30 May 2025).

Reddy, S. S., Gadiraju, M., & Maheswara Rao, V. (2022). Analyzing student reviews on teacher performance using long short-term memory. In *Innovative data communication technologies and application: Proceedings of ICIDCA 2021* (pp. 539–553). Springer.

Rodríguez Velasco, C. L., García Villena, E., Brito Ballester, J., Durántez Prados, F. Á., Silva Alvarado, E. R., & Crespo Álvarez, J. (2023). Forecasting of post-graduate students' late dropout based on the optimal probability threshold adjustment technique for imbalanced data. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(4), 120–155. [CrossRef]

Sakho, A., Malherbe, E., & Scornet, E. (2024). Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants. *arXiv*, arXiv:2402.03819.

Salas-Pilco, S. Z., & Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: A systematic review. *International Journal of Educational Technology in Higher Education*, *19*(1), 21. [CrossRef]

Salas-Pilco, S. Z., Yang, Y., & Zhang, Z. (2022). Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review. *British Journal of Educational Technology*, *53*(3), 593–619. [CrossRef]

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*, 100003. [CrossRef]

Shen, Y., Lai, E. M., & Mohaghegh, M. (2020). The role of RNNs for contextual representations: A case study using DMN-plus. In *Proceedings of the 2020 4th international symposium on computer science and intelligent control* (pp. 1–5). Association for Computing Machinery.

Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 345–356). Association for Computing Machinery.

Tian, X., Tang, S., Zhu, H., & Xia, D. (2022). Real-time sentiment analysis of students based on mini-Xception architecture for wisdom classroom. *Concurrency and Computation: Practice and Experience*, *34*, e7059. [CrossRef]

Toçoğlu, M. A., & Onan, A. (2021). Sentiment analysis on students' evaluation of higher educational institutions. In *Intelligent and fuzzy techniques: Smart and innovative solutions: Proceedings of the INFUS 2020 conference, Istanbul, Turkey, 21–23 July 2020* (pp. 1693–1700). Springer.

Viveka, M., & Priya, N. S. (2024). An efficient hybrid deep learning framework for predicting student academic performance. *Salud, Ciencia y Tecnología-Serie de Conferencias*, *3*, 759. [CrossRef]

Wahyuddin, E. P., Caraka, R. E., Kurniawan, R., Caesarendra, W., Gio, P. U., & Pardamean, B. (2025). Improved LSTM hyperparameters alongside sentiment walk-forward validation for time series prediction. *Journal of Open Innovation: Technology, Market, and Complexity*, *11*(1), 100458. [CrossRef]

Yi, W., Huang, X., Kuzmin, S., Gerasimov, I., & Luo, Y. (2025). Seekg: Sentiment analysis for E-Learning evaluation incorporating knowledge graphs. *Education and Information Technologies*, *30*, 16291–16320. [CrossRef]

https://doi.org/10.3390/educsci16030399