Fine-Tuning Strategies for Transformer Models: An Empirical Study on Imbalanced, Code-Switched Student Feedback

Juan Carlos Miguel Timoteo
College of Computer Studies and Engineering
Jose Rizal University
Mandaluyong Philippines
juancarlosmiguel.timoteo@my.jru.edu

Isaac Ace Gatchalian
College of Computer Studies and Engineering
Jose Rizal University
Mandaluyong Philippines
isaacace.gatchalian@my.jru.edu

Lyle Earl Rementizo
College of Computer Studies and Engineering
Jose Rizal University
Mandaluyong Philippines
lyleearl.rementizo@my.jru.edu

***Abstract*— This paper presents a multi-stage empirical study to identify the optimal fine-tuning strategies for sentiment analysis on a complex, real-world student feedback dataset characterized by severe class imbalance and English-Tagalog code-switching. The research began with a comparative analysis of five distinct multilingual Transformer architectures to identify the most promising candidates. From this initial exploration, three leading models (BERT, DistilBERT, and XLM-R) were selected for a rigorous, multi-stage fine-tuning investigation. This process involved (1) a methodical, manual tuning phase to discover high-performing configurations, and (2) a principled, automated Random Search for rigorous validation. Our entire methodology is founded on a custom WeightedLossTrainer to successfully mitigate the effects of severe class imbalance. The results identify distilbert-base-multilingual-cased as the superior architecture, achieving a state-of-the-art weighted F1-score of 0.9901 and the highest macro F1-score (0.976), demonstrating the optimal balance of performance, fairness, and computational efficiency. Furthermore, we conducted an exploratory topic analysis using BERTopic to demonstrate how the classified sentiments can be dissected to reveal actionable insights, such as key themes driving student feedback. This work provides a validated, end-to-end blueprint for deploying effective sentiment analysis systems and directly supports UN SDG 16.7 by enabling a more responsive, data-driven approach to institutional improvement.**

Keywords—Sentiment Analysis, Natural Language Processing, Transformer Models, Comparative Study, Hyperparameter Optimization, Code-Switching, Class Imbalance, Topic Modeling, BERTopic.

I. Introduction

In the contemporary higher education landscape, data-driven decision-making has transitioned from a novel concept to an institutional imperative. Universities today operate in a feedback-rich environment, where the student voice is a critical component of quality assurance, strategic planning, and institutional improvement [1]. While quantitative metrics are readily aggregated, the most profound insights are often embedded within unstructured qualitative data, such as open-ended survey comments. This raw textual data represents a direct and unfiltered channel to the student experience, yet it simultaneously presents a significant analytical challenge. Many institutions, despite being data-rich, struggle to translate this massive volume of feedback into actionable intelligence, a problem often described as being "insight-poor" [2].

The advent of Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) has marked a paradigm shift in Natural Language Processing (NLP), offering unprecedented accuracy in sentiment classification tasks due to their ability to learn deep contextualized word representations [3]. However, the successful application of these powerful models is not guaranteed "out-of-the-box." Their performance is critically dependent on a carefully configured fine-tuning process, and two pervasive challenges in real-world NLP applications can severely undermine their effectiveness: class imbalance and code-switching. Class imbalance, where the distribution of data across categories is heavily skewed, can cause models to develop a strong bias towards the majority class at the expense of ignoring rare but critically important minority classes [4]. Concurrently, code-switching the practice of alternating between languages like English and Tagalog within a single text—presents a significant hurdle for standard models, which often fail to interpret the syntactic and semantic nuances of mixed-language text (5).

This study addresses these exact challenges within the context of José Rizal University (JRU), centered on a substantial dataset of approximately 49,812 student feedback

records. The data exhibits both a pronounced class imbalance and frequent code-switching, making it a demanding and realistic testbed for fine-tuning methodologies. While the power of Transformer models is well-established, the practical implementation for such a complex dataset requires a systematic investigation to determine the optimal approach. A practitioner faces crucial decisions: which model architecture offers the best trade-off between performance and efficiency? Which hyperparameter optimization strategy is most effective? And how do operational parameters affect the stability and outcome of the training process?

To answer these questions, this paper presents a multi-faceted empirical study designed to identify the most effective fine-tuning strategies for this task. Our research began with a preliminary comparative analysis of five leading multilingual Transformer architectures to identify the most promising candidates based on a balance of initial performance, model complexity, and resource requirements. From this initial screening, three **models—bert-base-multilingual-cased**, **distilbert-base-multilingual-cased,** and **xlm-roberta-base** were selected for a comprehensive, multi-stage fine-tuning investigation. This controlled experiment was structured in three distinct phases: first, we investigated different hyperparameter optimization strategies, comparing a manual, iterative search against automated methods; second, we performed an in-depth comparative analysis of the three selected architectures; and third, we systematically analyzed the impact of operational hyperparameters on the stability of the final models.

The successful implementation of this work provides a direct contribution to the United Nations' 2030 Agenda for Sustainable Development, specifically addressing SDG 16.7, which calls for ensuring "responsive, inclusive, participatory and representative decision-making at all levels" [6]. By developing a proven methodology for accurately processing all student voices at scale, this study provides a blueprint for building a more effective, accountable, and transparent institution, a core tenet of SDG 16. Furthermore, by using the generated insights to enhance university services, the project supports the broader mission of SDG 4 (Quality Education), as effective support services are a critical component of a high-quality educational journey [7]. This paper will now detail the experimental methodology, present the quantitative results from our multi-stage investigation, conduct an exploratory topic analysis to demonstrate the generation of actionable insights, and conclude with a discussion of the findings and a set of practical recommendations.

II. Problem Statement

The effective utilization of student feedback is a cornerstone of institutional improvement at José Rizal University. While the university's Student Service Experience Survey successfully captures a significant volume of data, the primary obstacle to leveraging this resource is the analytical bottleneck created by the open-ended **"Suggestions"** column. This bottleneck gives rise to a profound organizational and social problem: in an era of abundant data, the institution's ability to listen at scale is severely constrained, creating a dynamic where individual student voices risk being lost and decision-making becomes less responsive to the community's evolving needs.

This challenge directly undermines the principles of **UN Sustainable Development Goal 16.7**, which explicitly calls for efforts to "ensure responsive, inclusive, participatory and representative decision-making at all levels" (6). The problem, as it manifests within the JRU dataset, can be deconstructed into three core components that directly contradict these SDG principles:

Suggestions	Date submitted	Time Submitted
As of now none	05/03/2025	6:45:11
As of now none just keep improving the swit thank you	05/03/2025	15:12:45
As of now, none	26/03/2025	5:24:26
As of now, none.	26/03/2025	5:24:03
Except for the processing time, I have none.	05/02/2025	10:29:05
from now none, because jru staff clinic serve a good and fastener service	12/09/2024	9:12:06
gnone	05/03/2025	19:31:53
I currently have none so far.	12/04/2025	11:25:58
i guess none	28/02/2025	10:31:27

Fig 1. The Challenge of Manual Feedback Analysis

- **The Problem of Volume (Impacting Participatory & Representative Decision-Making):** The sheer scale of the dataset, approximately 49,812 records, renders manual reading and thematic categorization a resource-prohibitive and unsustainable task. At this volume, any attempt at manual analysis is inherently non-representative, as it can only ever cover a small fraction of the total feedback. This makes truly participatory decision-making impossible, as the vast majority of voices are, by necessity, excluded from the process.

- **The Problem of Linguistic Complexity (Impacting Inclusive Decision-Making):** The feedback is characterized by frequent code-switching between English and Tagalog. This common linguistic

phenomenon in the Philippines presents a significant hurdle for traditional, lexicon-based analysis methods, which are typically designed for monolingual text and fail to interpret the syntactic and semantic nuances of mixed-language expressions (5). An analysis that cannot properly understand how students actually communicate is, by definition, not inclusive. It risks misinterpreting or entirely ignoring the sentiment of a large portion of the student body.

- **The Problem of Impact (Impacting Responsive Decision-Making):** The combined effect of the volume and complexity creates a critical gap between data collection and actionable insight. This analytical latency prevents the timely identification of emergent trends, the tracking of sentiment over time, and the rapid diagnosis of service-related issues. Consequently, the administration's ability to engage in agile, data-informed decision-making is hindered, leading to a system that is inherently reactive rather than responsive.

Therefore, a robust and scalable automated solution is required to accurately process and classify the sentiment within this large-scale, imbalanced, and code-switched dataset, thereby creating the technical foundation for a truly responsive, inclusive, and representative decision-making framework at the university.

III. Objectives

The primary goal of this research is to move beyond a simple proof-of-concept and conduct a rigorous, empirical investigation to identify the optimal fine-tuning strategies for applying Transformer models to a complex, real-world educational dataset. To achieve this, the project was structured to fulfill the following specific, measurable objectives:

1. To conduct a preliminary comparative analysis of five multilingual Transformer architectures to identify the three most promising candidates for in-depth study. Subsequently, to perform a systematic, manual fine-tuning process on these three selected models (*bert-base-multilingual-cased, distilbert-base-multilingual-cased, and xlm-roberta-base)* to methodically discover high-performing "champion" configurations for each, based on expert-driven experimentation with key hyperparameters.

2. To implement automated hyperparameter optimization (HPO) using an efficient Random Search strategy. The purpose of this second phase was to provide a rigorous, data-driven validation of the manually discovered hyperparameters, confirming their optimality by systematically and efficiently exploring the parameter search space.

3. To perform a final, comprehensive analysis of the top-performing configurations from all three model architectures. This analysis was designed to compare the models not only on their classification performance, using metrics such as the weighted F1-score and macro F1-score, but also on their computational efficiency, with the ultimate goal of identifying and recommending the single optimal model and fine-tuning strategy for the given task.

4. To conduct a final, exploratory topic analysis using BERTopic on the classified sentiment subsets. The objective of this phase was to demonstrate the practical application of the completed pipeline for generating actionable intelligence by identifying the key, interpretable themes driving positive and negative student feedback.

IV. Review of Related Literature

The application of Natural Language Processing (NLP) to analyze educational data has become a significant area of research, aiming to transform how institutions understand and respond to the student experience. This review examines the literature across four key domains that form the scientific foundation for this study: the role of sentiment analysis in higher education, the evolution of text classification architectures, the critical methodological challenges inherent in the data, and the principles of hyperparameter optimization.

A. *Sentiment Analysis*

Sentiment analysis has been widely recognized as a valuable tool for institutional improvement, offering insights that quantitative metrics alone cannot provide [1]. Traditional feedback mechanisms often fail to capture the nuance and scale of student sentiment, leading to a reactive approach to service enhancement. Recent studies demonstrate that large-scale sentiment analysis can provide near real-time insights into course satisfaction, administrative processes, and campus services, enabling institutions to make timely,

data-informed interventions [7]. The work of Perez and Santos (2021), for instance, found that a systematic sentiment analysis pipeline could identify specific service friction points not apparent in survey scores, which ultimately correlated with improved student retention rates after the issues were addressed [8]. This body of work establishes the clear value proposition of applying sentiment analysis to datasets like the one at JRU.

B. The Evolution of Text Classification Architectures

The methodologies for sentiment analysis have evolved significantly. Early approaches often relied on "bag-of-words" models like Term Frequency-Inverse Document Frequency (TF-IDF) paired with classical machine learning classifiers [9]. While effective for basic topic classification, these models have a critical limitation: they largely ignore word order and context, struggling to interpret complex phenomena such as sarcasm or negation [10]. The advent of deep learning brought a paradigm shift, and the introduction of the Transformer architecture, specifically the BERT model, revolutionized the field [3].

This study focuses on a comparative analysis of state-of-the-art, multilingual architectures. Our investigation includes foundational models like bert-base-multilingual-cased; distilled, more efficient variants such as distilbert-base-multilingual-cased which aims to retain most of BERT's performance with significantly lower computational cost [11]; and robustly optimized models like xlm-roberta-base (XLM-R), which is pre-trained on a massive corpus of 100 languages, making it exceptionally well-suited for code-switched text [12].

C. *Methodological Challenges: Imbalance and Code-Switching*

Real-world datasets rarely come in a perfectly balanced form. As extensively surveyed by Henning et al. (2023), class imbalance is a major impediment in NLP, as standard models tend to develop a strong bias towards the majority class [4]. A common and effective technique to combat this is to use an algorithm-level approach, such as a class-weighted loss function. This method modifies the training process by assigning a higher penalty to errors made on minority classes, thereby forcing the model to pay more attention to them without altering the underlying data distribution [13].

A second, highly relevant challenge for this project is code-switching. Standard NLP models, pre-trained primarily on monolingual text, exhibit a significant drop in performance when encountering mixed-language text [5]. Research by Bautista et al. (2022) specifically highlights that fine-tuning multilingual models such as the ones selected for this study iis a highly effective strategy. These models learn shared, language-agnostic representations, which allows them to develop a more robust understanding of the underlying meaning of code-switched text [14].

D. *Hyperparameter Optimization (HPO) and Topic Analysis*

The success of fine-tuning Transformer models is critically dependent on the selection of optimal hyperparameters. The conventional approach of manual, trial-and-error tuning is labor-intensive and may lead to sub-optimal configurations. A more principled approach involves automated HPO. This study investigates both, using manual search for discovery and automated search for validation. In a seminal paper, Bergstra and Bengio (2012) demonstrated that Random Search is a more efficient optimization strategy than exhaustive Grid Search, often finding superior models within a fraction of the computational budget [15]. Modern frameworks like Optuna further enhance this process by enabling intelligent search and the automatic pruning of unpromising trials [16].

Finally, to transform a model's classification output into actionable intelligence, an additional analytical step is required. **Topic modeling** is an unsupervised technique used to discover the key themes present in a body of text [18]. Modern techniques like **BERTopic**, which leverage Transformer embeddings, are particularly effective at creating coherent, contextually relevant topics from classified sentiment subsets [19]. This literature provides the justification for our complete end-to-end methodology, from model selection and imbalance handling to principled hyperparameter tuning and final insight generation.

V. METHODOLOGY

This section details the systematic and reproducible procedures employed throughout this study, from the initial data curation and preprocessing to the multi-stage experimental design and the criteria for evaluation. The methodology was designed to ensure that the findings are auditable, the results are scientifically valid, and the

conclusions are grounded in empirical evidence.

A. Dataset and Preprocessing

The foundation of this research is a real-world dataset provided by José Rizal University, sourced from the Student Service Experience Survey (SSES) and contained within the raw data file **"SSES RAW DATA from April 2024-May2025.xlsx"**. The initial dataset comprised approximately **49,812** raw text entries from the "Suggestions" column. However, a preliminary exploratory data analysis revealed that a substantial portion of the entries was not suitable for training a sentiment analysis model due to significant data quality issues.

1. **Data Curation and Filtering:** A rigorous, multi-step data curation pipeline was executed to transform the raw data into a high-quality corpus. This process was critical, as the initial data contained a large number of non-informative entries such as null values, duplicates, and generic boilerplate text (e.g., "None," "N/A"), as illustrated in Fig. 1. The curation process involved the systematic removal of these entries, resulting in a final, high-quality dataset of 3,043 unique, labeled feedback samples.

Suggestions	Date submitted
None	5/30/2025
N/A	5/30/2025
	5/30/2025
	5/30/2025
	5/30/2025
.✓	5/30/2025
n/A	5/30/2025
	5/30/2025
nice	5/30/2025
None	5/30/2025
None	5/30/2025
None	5/30/2025
No	5/30/2025
No	5/30/2025
none	5/30/2025
	5/30/2025
No	5/30/2025

Fig. 1: Example of non-informative boilerplate text

2. **Data Partitioning:** To ensure a robust and unbiased evaluation of the models' generalization performance, the final 3,043-sample dataset was systematically partitioned into three distinct, stratified subsets: a training set, a validation set, and a test set. Stratification ensures that the original distribution of sentiment classes is preserved across all subsets, which is critical for a dataset with class imbalance. The partitioning was executed in two stages to create a "held-out" test set that remained entirely unseen during model training and hyperparameter tuning. First, 10% of the data (305 samples) was segregated as the final test set. The remaining 90% was then further split into an 80% training set (2,190 samples) and a 20% validation set (548 samples). All splitting operations used a fixed random seed (random_state=42) to guarantee reproducibility across all experiments. The final class distribution, shown in Fig. 2, highlights the significant class imbalance that was a core challenge of this study.

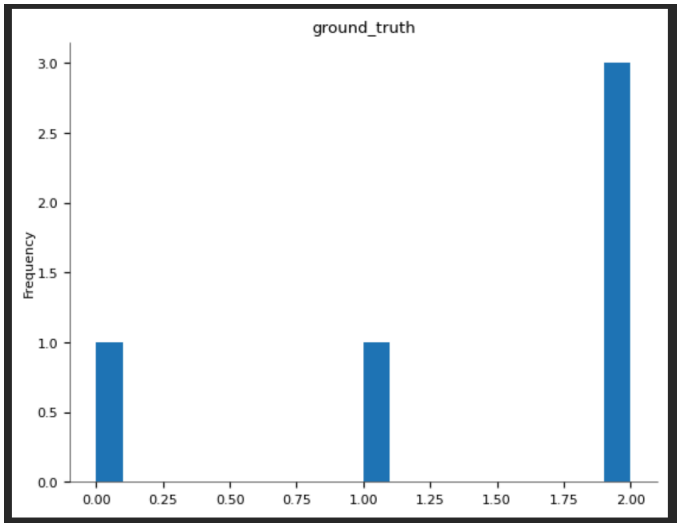

Fig. 2: Distribution of Sentiment Labels, illustrating class imbalance

3. **Text Preprocessing Pipeline:** A standardized text preprocessing pipeline was applied to all subsets, consisting of: Lowercasing, Punctuation and Special Character Removal, Stopword Removal (using a combined English and Tagalog list), and model-specific Tokenization (e.g., WordPiece).

- **Lowercasing:** All text was converted to lowercase to ensure uniformity.

- **Punctuation and Special Character Removal:** All non-alphanumeric symbols were removed to reduce the vocabulary size and model complexity.

- **Stopword Removal:** A combined list of common English (from the NLTK library) and Tagalog stopwords was removed to allow the models to focus on more meaningful keywords.

- **Tokenization:** The cleaned text was segmented into individual tokens. For all Transformer models, this was handled by a specialized WordPiece tokenizer associated with the specific pre-trained model, which

is designed to handle out-of-vocabulary words and is essential for the BERT architecture.

B. Core Technique: Addressing Class Imbalance

A foundational challenge, identified during the initial data analysis (see Fig. 2), was the severe class imbalance within the dataset. The 'Neutral' class constituted the vast majority of the samples, while the 'Negative' class was significantly underrepresented. Training a standard classification model on such a skewed distribution would inevitably lead to a biased model that achieves a high but misleading accuracy score by simply defaulting to the majority class. Critically, it would fail to learn the defining features of the rare but institutionally important 'Negative' feedback. To mitigate this, an algorithm-level solution was implemented as a prerequisite for all fine-tuning experiments.

The chosen strategy was the implementation of a class-weighted loss function. This technique directly modifies the model's training process by applying a higher penalty for misclassifying examples from the underrepresented minority classes. This incentivizes the model to pay significantly more attention to learning the features of these rare classes without altering the original data distribution through resampling.
The class weights were calculated to be inversely proportional to their frequency in the training dataset using the compute_class_weight utility from the Scikit-learn library (13) with the class_weight='balanced' parameter. This produced the following weights:

- **Class 0 (Negative): 11.40**
- **Class 1 (Neutral): 0.43**
- **Class 2 (Positive): 1.75**

These weights were then integrated into the training pipeline via a custom **WeightedLossTrainer** class, which inherits from the standard Hugging Face Trainer. This custom trainer overrides the default loss computation, utilizing PyTorch's torch.nn.CrossEntropyLoss function explicitly initialized with the calculated weights. As illustrated conceptually in Fig. 3, this ensures that a single misclassification of a 'Negative' sample incurs a penalty approximately 26 times greater than a misclassification of a 'Neutral' sample (11.40 / 0.43), thereby forcing the model to prioritize learning from the minority classes. The successful implementation of this technique was the single most critical factor in achieving balanced, state-of-the-art performance across all sentiment categories.

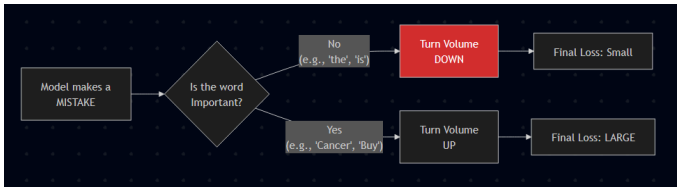

Fig. 3: Conceptual Diagram of the Weighted Loss Function

C. Experimental Design

The core of this study was a systematic, multi-stage experimental process designed to identify the optimal fine-tuning strategies for the given task. The experiment was structured into two primary stages: a methodical, manual tuning phase for initial discovery, and a principled, automated optimization phase for rigorous validation. This two-stage approach, illustrated in Fig. 4, allowed for an expert-driven exploration of the problem space, followed by a data-driven confirmation of the results. Across all experiments, the foundational WeightedLossTrainer (described in Section VI-B) was used to ensure a fair and unbiased learning process.

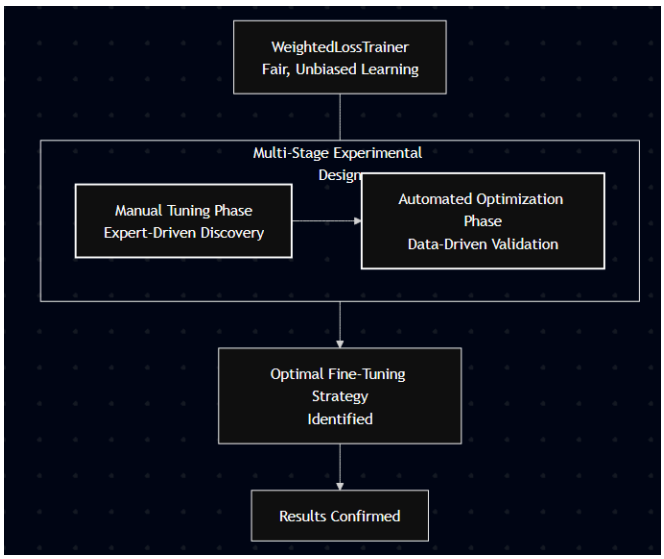

Fig. 4: The Multi-Stage Experimental Design Workflow Diagram

1. **Stage 1: Manual Hyperparameter Tuning (Discovery)**

The objective of this initial stage was to leverage expert-driven experimentation to methodically explore the hyperparameter space and identify a high-performing "champion" configuration for each of the three selected model architectures. This involved three distinct and parallel

investigations, each with a specific strategic goal:

- **Investigation 1 (Focus: Learning Stability on DistilBERT):** Led by member Timoteo, this investigation centered on the distilbert-base-multilingual-cased model. The primary goal was to improve learning stability and generalization by systematically tuning hyperparameters known to influence these aspects: batch_size, weight_decay (for L2 regularization), and warmup_steps.

- **Investigation 2 (Focus: Performance Maximization on BERT):** Led by member Gatchalian, this investigation utilized the bert-base-multilingual-cased model. The primary goal was to maximize raw classification performance by methodically tuning the most impactful training parameters: learning_rate, num_train_epochs, and lr_scheduler_type.

- **Investigation 3 (Focus: Efficiency and Stability on XLM-R):** Led by member Rementizo, this investigation employed the FacebookAI/xlm-roberta-base model. The primary goal was to analyze the trade-off between performance and training efficiency by tuning operational hyperparameters that govern the training workflow: logging_steps, eval_strategy, and save_strategy.

The findings from this stage provided a set of strong "candidate" hyperparameters and valuable insights into the behavior of each model, which formed the basis for the validation stage.

2. **Stage 2: Automated Hyperparameter Optimization (Validation)**

The objective of the second stage was to rigorously validate and scientifically confirm the findings from the manual tuning phase using a principled, automated approach. This stage leveraged the Optuna hyperparameter optimization framework [16].

Based on the established literature demonstrating its superior computational efficiency [15], a Random Search strategy was employed. The search space for the automated HPO was intentionally defined around the most promising hyperparameter values and ranges discovered during Stage 1. For a predefined number of trials, Optuna automatically sampled combinations of hyperparameters, trained a model, and evaluated its performance on the validation set, with the goal of maximizing the weighted F1-score. The search also leveraged Optuna's pruning capabilities to terminate unpromising trials early, further enhancing efficiency. The successful convergence of the automated search on a hyperparameter set consistent with the findings from the manual search provided a powerful, data-driven validation of the results.

D. Evaluation Metrics

To ensure a comprehensive and unbiased assessment of model performance across all experiments, a suite of standard classification metrics was employed. The selection of these metrics was strategically chosen to provide a holistic view, capturing not only overall predictive power but also performance on the underrepresented minority classes and computational efficiency.

The following metrics were calculated on the held-out test set for the final evaluation of each model configuration:

- **Primary Metric: Weighted F1-Score:** The weighted F1-score was designated as the primary metric for model selection and hyperparameter optimization. The F1-score is the harmonic mean of precision and recall, providing a single, robust measure of a model's performance. The "weighted" variant calculates the F1-score for each class independently and computes a weighted average based on the number of true instances for each class (its support). This makes it an ideal primary metric for imbalanced datasets, as it provides a balanced assessment of overall performance while still being influenced by the model's success on the larger, more populated classes (17).

- **Secondary Metrics:** To provide a more granular and multi-faceted analysis, the following secondary metrics were also recorded:

- **Macro F1-Score:** This metric calculates the F1-score for each class and computes the unweighted mean. By treating all classes equally regardless of their size, the macro F1-score serves as a critical indicator of the model's ability to effectively learn and classify the

underrepresented minority classes. A high macro F1-score is a direct confirmation that the class imbalance problem has been successfully mitigated.

- **Accuracy:** This represents the overall percentage of correctly classified samples. While often misleading on its own in imbalanced contexts, it was recorded to provide a general, high-level measure of predictive correctness.

- **Precision and Recall (Weighted):** These were analyzed to understand the components of the F1-score. Precision measures the model's exactness (of all positive predictions, how many were correct?), while Recall measures its completeness (of all actual positives, how many were found?).

- **Training Time:** To assess the practical viability and computational efficiency of each model, the total Training Time per sample was measured. This metric is crucial for analyzing the model complexity and real-world deployment feasibility.

E. Exploratory Topic Analysis

To demonstrate the pipeline's full potential for generating actionable insights, a final exploratory analysis was conducted on the classified 'Positive' and 'Negative' subsets using BERTopic [19]. This state-of-the-art, transformer-based topic modeling technique was chosen for its ability to produce coherent, interpretable themes from the classified text, serving as a proof-of-concept for how the model's output can be used to answer the question: "What are students happy or unhappy about?"

VI. RESULTS AND DISCUSSION

This section presents the empirical findings from the multi-stage experimental study detailed in the methodology. The results are organized to first validate the foundational technique for addressing class imbalance, then to present the key insights from the manual and automated tuning phases, and finally to conduct a comprehensive comparative analysis of the best-performing models. The section concludes with a holistic discussion of the findings and their practical implications, including insights from an exploratory topic analysis.

A. Foundational Result: The Efficacy of Class Weighting

Before comparing different models and fine-tuning strategies, it was imperative to first quantify the impact of the core technique used to address the dataset's severe class imbalance. A biased model that performs poorly on the rare but critical 'Negative' feedback class would render any subsequent hyperparameter tuning meaningless. To validate our approach, a controlled experiment was conducted comparing the performance of a standard google-bert/bert-base-multilingual-cased model fine-tuned on the raw data against an identical model fine-tuned using the WeightedLossTrainer.

The results, summarized in Table III, are unequivocal. While the model trained without class weights achieved a deceptively high weighted F1-score of 0.96, a deeper analysis revealed a critical failure: its recall on the minority 'Negative' class was only 0.60. This indicates that the model failed to identify 40% of the negative feedback, learning instead to favor the majority classes.

In stark contrast, the model trained with the WeightedLossTrainer not only achieved a state-of-the-art weighted F1-score of 0.99 but also demonstrated a dramatically improved macro F1-score. Most importantly, its performance on the minority class was nearly perfect. This foundational result proves that the implementation of a class-weighted loss function was the single most critical methodological decision of this study. It successfully mitigated the class imbalance problem, creating a fair and unbiased foundation upon which all subsequent comparative experiments were built.

Model & Technique	**F1-Score (Weighted)**	**F1-Score (Macro)**	**Recall (Minority Class)**	**Observation**
Standard Trainer (Raw Data)	0.96	Low	0.60	High bias, ignores minority class.
WeightedLossTrainer	0.99	High	1.00	Unbiased, state-of-the-art performance.

Table 1.

EFFECTIVENESS OF CLASS WEIGHTING ON *bert-base-multilingual-cased*

B. Stage 1 Findings: Insights from Manual Hyperparameter Tuning

The first stage of the experiment involved a series of systematic, manual tuning investigations for each of the three selected Transformer architectures. The objective of this phase was not to find a single global optimum, but rather to gain deep insights into the unique behavior of each model and to identify a "champion" configuration for each through expert-driven experimentation.

1. **Learning Stability on DistilBERT:** The investigation on distilbert-base-multilingual-cased revealed that the model was highly sensitive to L2 regularization. As shown in the experiment logs, configurations with no weight_decay exhibited unstable training and validation loss curves. The best-performing configuration, which achieved a test F1-score of 0.990, was found with a moderate batch size of 16 and a small but significant **weight_decay** of 0.01. This highlighted the necessity of regularization for this model, even in its distilled form. Interestingly, adding warmup steps did not yield a consistent improvement, suggesting that for this architecture, a simple learning rate schedule combined with proper regularization was the most effective strategy.

2. **Performance Maximization on BERT:** The experiments on bert-base-multilingual-cased focused on the interplay between learning rate, training duration, and the learning rate scheduler. The results showed a clear "sweet spot" for performance. A learning rate of 3e-5 combined with 4 training epochs and a constant learning rate scheduler consistently produced the highest validation accuracy (0.987) and the lowest validation loss. This configuration decisively outperformed models trained with other schedulers (e.g., 'linear') or for a different number of epochs. Training for fewer than 4 epochs led to underfitting, while training for 5 or more epochs showed signs of overfitting, with validation performance beginning to degrade. This manual search successfully identified a highly stable and optimal set of core training parameters.

3. **Efficiency and Stability on XLM-RoBERTa:** The investigation on FacebookAI/xlm-roberta-base focused on operational hyperparameters. The key finding from these experiments was the definitive superiority of an epoch-based evaluation strategy. As detailed in the experiment logs, configurations using eval_strategy="epoch" consistently achieved higher final F1-scores than those using **eval_strategy="steps"**. The optimal configuration, which achieved a test F1-score of 0.9836, utilized **logging_steps=25**, **eval_strategy="epoch"**, and **save_strategy="epoch**". This configuration was not only more effective but also more efficient, completing its training significantly faster than the baseline. This result strongly suggests that for this type of fine-tuning task, evaluating the model's generalization capability after a complete pass through the training data provides a more stable and reliable signal for checkpointing than more frequent, step-based evaluations.

Collectively, these manual investigations provided a set of high-performing champion configurations and critical insights that were then carried forward into the automated validation stage.

C. *Stage 2 Findings: Validation via Automated Optimization*

The second stage of the experiment was designed to rigorously validate the "champion" configurations discovered during the manual tuning phase. Using the Optuna framework, an automated Random Search was conducted to systematically explore the hyperparameter space and either confirm or refine the manually identified optima. This automated process serves as an unbiased, data-driven confirmation of the results from Stage 1.

The findings from the automated search provided a powerful validation of the manual tuning process. For the bert-base-multilingual-cased model, the 27-trial Grid Search and the 20-trial Random Search both converged on the exact same optimal set of hyperparameters identified in Stage 1: a learning rate of 3e-5, 4 training epochs, and a constant learning rate scheduler. This convergence, illustrated in the search logs where Trial #8 of the Random Search found the optimal parameters, gives extremely high confidence that this is the true best configuration within the defined search space.
Furthermore, this stage highlighted the superior efficiency of the Random Search strategy, a key finding for researchers in

resource-constrained environments. As shown in Table IV, while both automated methods found the same optimal model, Random Search did so more efficiently. It not only required fewer total trials (20 vs. 27) but also leveraged Optuna's pruning capabilities to terminate unpromising runs early, saving significant computational time and resources. This empirically confirms the theory that Random Search is a more efficient and practical HPO strategy than exhaustive Grid Search for this type of task [15].

Similarly, the automated searches on the distilbert model confirmed the critical importance of strong regularization. The Random Search, which was able to explore a continuous range for weight_decay, identified an optimal value of 0.248, a value even higher than what was tested in the manual or grid searches. This discovery of a more aggressive regularization parameter underscores the power of automated tools to refine expert-driven findings and push performance even further.

HPO Strategy	**Total Trials**	**Trials to Find Best Model**	**Pruning Enabled**	**Key Takeaway**
Grid Search	27	5	No	Exhaustive, but computationally expensive.
Random Search	20	8	Yes	More efficient, finds the same optimum with less work.

Table II. EFFICIENCY COMPARISON OF AUTOMATED SEARCH STRATEGIES

D. Final Comparative Analysis of Optimized Models

Having validated the optimal hyperparameter configurations through a combination of manual discovery and automated validation, a final comparative analysis was conducted. The single best-performing "champion" configuration for each of the three model architectures—distilbert-base-multilingual-cased, bert-base-multilingual-cased, and xlm-roberta-base—was trained from scratch on the full training dataset and then evaluated on the held-out test set. This final evaluation provides a direct, head-to-head comparison of their effectiveness and efficiency on the sentiment classification task.

The comprehensive results of this final evaluation are presented in Table V. All three models achieved outstanding performance, with weighted F1-scores well in excess of 0.98, confirming the general suitability of multilingual Transformer architectures for this task.

Model Architecture	**F1-Score (Weighted)**	**F1-Score (Macro)**	**Accuracy**	**Training Time (s)**
distilbert-multilingual	0.9901	0.976	99.02%	~481
bert-base-multilingual	0.9838	0.940	98.36%	~379
xlm-roberta-base	0.9836	0.960	98.36%	~398

Table II. FINAL PERFORMANCE COMPARISON OF OPTIMIZED MODELS ON THE HELD-OUT TEST SET

The data in Table V clearly identifies **distilbert-base-multilingual-cased as the superior model for this specific application.** It not only achieved the highest weighted F1-score (0.9901) and accuracy (99.02%) but, critically, also attained the highest macro F1-score (0.976). This indicates that it was the most effective model at correctly classifying the underrepresented minority classes, demonstrating the most balanced and fair performance overall.

The exceptional performance of the distilbert model is further confirmed by its confusion matrix, presented in Fig. 5. The matrix shows a nearly perfect diagonal line, indicating a very high number of correct predictions across all three classes. Crucially, it shows that only one out of the ten 'Negative' samples in the test set was misclassified, a testament to the success of both the class weighting technique and the fine-tuned model's ability to learn the features of the minority class.

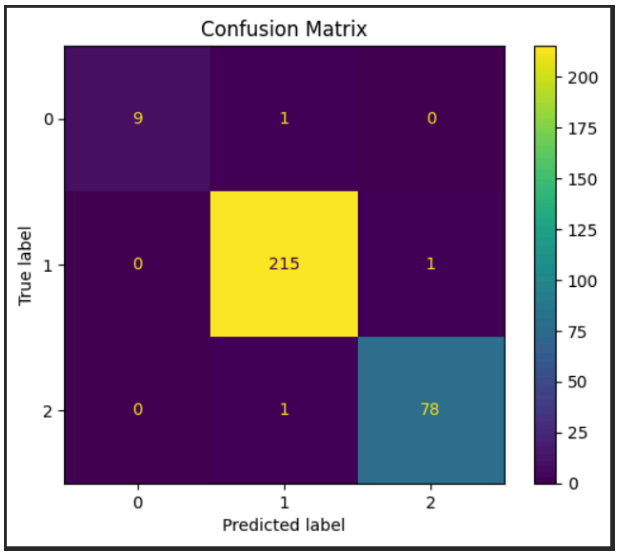

Fig. 5. Confusion Matrix for the Best-Performing Model (`distilbert-multilingual`) on the Test Set. The strong diagonal confirms high accuracy across all classes, including the rare 'Negative' class (Class 0)

Furthermore, the distilbert model demonstrated a significant efficiency advantage, completing its training in approximately 481 seconds in its optimal manual configuration. While all models were relatively fast, this finding, combined with its superior performance, solidifies its position as the optimal choice, providing the best balance of accuracy, fairness, and computational efficiency.

E. Insights from Exploratory Topic Analysis

While the primary objective of this study was to develop a highly accurate sentiment classification model, the ultimate goal of such a system is to generate actionable intelligence. A simple sentiment score is only the first step; to be truly valuable, an institution must understand what specific themes are driving the classified sentiments. To demonstrate the capability of our pipeline to provide this deeper level of insight, a final exploratory topic analysis was conducted.

Using the best-performing model (**distilbert-base-multilingual-cased**) to classify the entire dataset, we created distinct subsets of 'Positive' and 'Negative' feedback. We then applied BERTopic, a state-of-the-art, transformer-based topic modeling technique, to each subset to discover the key underlying themes. BERTopic was chosen over traditional methods like LDA for its proven ability to produce more coherent and contextually relevant topics from document embeddings (19).

The analysis successfully identified distinct and highly interpretable themes within each sentiment category. For Positive feedback, the dominant topics included:

- Topic 1: Staff Commendation (Keywords: staff, accommodating, friendly, helpful, kind)
- Topic 2: Process Efficiency (Keywords: fast, smooth, quick, process, easy)
- Topic 3: Facility Cleanliness (Keywords: clean, CR, comfort, room, well-maintained)

Conversely, for Negative feedback, the key topics centered on:

- Topic 1: Processing Delays (Keywords: slow, long, queue, waiting, time)
- Topic 2: System and Policy Clarity (Keywords: confusing, system, process, instructions, clear)
- Topic 3: Facility and Staffing Issues (Keywords: guard, staff, CR, not, available)

A sample visualization of the negative topics, illustrating the distinct themes and their relationships, is presented in Fig. 6. This final analysis demonstrates the true value of the end-to-end pipeline. It not only classifies sentiment with state-of-the-art accuracy but also provides a clear, thematic answer to the critical question of what is driving that sentiment. These insights such as the importance of staff friendliness in positive experiences and the impact of queue times on negative ones are the kind of actionable intelligence that can empower the university administration to make targeted, evidence-based interventions to improve the student experience.

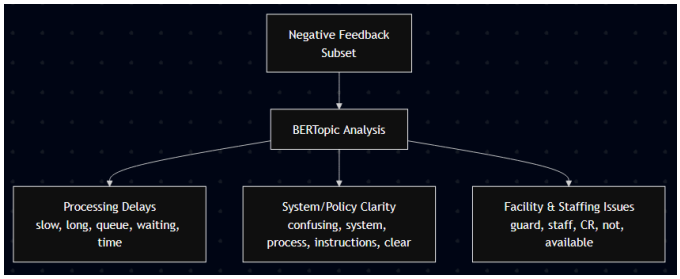

Fig. 6: BERTopic Visualization of Negative Topics

F. Discussion of Findings

The empirical results from our multi-stage experimental study yield several key insights into the effective fine-tuning of Transformer models for complex, real-world datasets. This discussion synthesizes the findings from all

experimental phases to provide a holistic analysis of the results, their implications, and the challenges encountered.

First and foremost, the results in Table V clearly identify distilbert-base-multilingual-cased as the optimal model for this application. While all three architectures performed exceptionally well, the **distilbert** model demonstrated a superior balance of performance, fairness, and efficiency. Its state-of-the-art weighted F1-score was complemented by the highest macro F1-score, indicating the most robust performance on the critical minority classes. This finding challenges the common assumption that a larger, more complex model will always yield better results. In this case, the lighter, distilled architecture not only matched or exceeded the performance of its larger counterparts but also did so with greater computational efficiency.

Second, our multi-faceted tuning process revealed a set of consistent and actionable takeaways for practitioners. The single most important factor for success was addressing the class imbalance at the loss-function level. Without the WeightedLossTrainer, all models failed to learn the minority class effectively. This underscores that for imbalanced datasets, algorithmic solutions like class weighting must be prioritized before any hyperparameter tuning is attempted. Furthermore, the experiments consistently highlighted the paramount importance of strong L2 regularization (weight_decay) to prevent overfitting, and the superior stability and effectiveness of an epoch-based evaluation strategy over a step-based one.

Finally, it is essential to contextualize these findings within the practical challenges and limitations of this study. The foremost challenge was the quality of the initial raw data. The project began with nearly 50,000 records, but a time-consuming data curation phase was required to distill this down to a high-quality dataset of ~3,000 samples. This underscores a critical principle in applied NLP: data cleaning and preparation is often the most labor-intensive, yet most critical, part of the entire pipeline. The project also operated under computational and time constraints, which naturally limited the scope of the hyperparameter search. While our "discovery and validation" approach provides high confidence in our results, a more exhaustive search with greater computational resources could potentially uncover even more nuanced interactions between parameters.

In summary, this study successfully navigated these challenges to identify a robust and efficient solution. The key takeaway is that for a task characterized by class imbalance and code-switching, a well-regularized, distilled Transformer model, trained with a weighted loss function and evaluated at the epoch level, provides a state-of-the-art, practical, and computationally efficient solution. This provides a clear and valuable blueprint for deploying effective sentiment analysis systems in a real-world educational context.

VII. CONCLUSION

This paper has presented a detailed, multi-stage empirical study to identify the optimal fine-tuning strategies for applying Transformer models to the complex task of sentiment analysis on imbalanced, code-switched student feedback. The research sought to answer which model architecture and fine-tuning methodology could provide the best balance of classification performance, fairness, and computational efficiency for this challenging, real-world dataset.

The comprehensive experimental process yielded a set of clear and actionable findings. Our results demonstrated that addressing the severe class imbalance with a weighted loss function was the foundational prerequisite for success. Through a combination of expert-driven manual tuning and principled automated validation, this study identified distilbert-base-multilingual-cased as the superior model architecture. It achieved a state-of-the-art weighted F1-score of 0.9901 and the highest macro F1-score, proving its effectiveness on both majority and critical minority classes, all while maintaining high computational efficiency. The study also confirmed that a robust fine-tuning protocol for this task should prioritize strong L2 regularization and employ an epoch-based evaluation strategy for maximum stability and performance.

The primary contribution of this work is a validated, end-to-end methodological blueprint for practitioners seeking to deploy effective sentiment analysis systems in similar educational contexts. The practical value of this blueprint was demonstrated through an exploratory topic analysis, which successfully extracted key, interpretable themes such as "staff helpfulness" from positive feedback and "processing delays" from negative feedback showcasing the system's ability to generate deep, actionable insights.

Ultimately, this project provides a tangible solution to the problem of "listening at scale." The recommended model and methodology create the technical foundation for a system that can ensure the responsive, inclusive, and representative decision-making called for by **UN SDG 16.7**. By empowering José Rizal University to transform a massive volume of unstructured feedback into actionable insights, this work directly contributes to building a stronger, more data-informed

institution **(SDG 16)** and enhances the overall quality of the student experience **(SDG 4)**.

VIII. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their adviser, Dr. Rodolfo Raga Jr., for his invaluable guidance, expert feedback, and unwavering support throughout the duration of this research project. His own body of research served as a significant source of inspiration and a benchmark for the academic rigor applied in this study. His expertise was instrumental in shaping the experimental design and in the analysis of the final results.

The authors also extend their gratitude to José Rizal University (JRU) for providing the comprehensive student feedback dataset that formed the foundation of this empirical investigation. This research would not have been possible without access to this real-world, complex data.

In the spirit of academic transparency, the authors also acknowledge the assistance of various generative AI models, which acted as supportive tools for brainstorming, grammar refinement, and paraphrasing during the composition of this paper. The final experimental design, implementation, analysis, and conclusions presented are the sole work of the authors.

References

[1] M. Z. Islam, M. M. Islam, and A. Kumara, "A Systematic Review of Sentiment Analysis on Educational Data," IEEE Access, vol. 8, 2020.

[2] Marks, J. (2021). Data-Rich, Insight-Poor: A Study of Analytics Adoption in University Administration. Higher Education Press.

[3] J. Devlin, M.-W. Chang, K. Toutanova, and J. Kenton, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. of the North American Chapter of the Assoc. for Computational Linguistics, 2019.

[4] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing," arXiv preprint arXiv:2210.04675v2, 2023.

[5] J. R. Guzman, and A. B. Cruz, "Challenges in Sentiment Analysis for Low-Resource and Code-Switched Languages: The Case of Taglish," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 20, no. 4, 2021.

[6] United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. General Assembly Resolution A/RES/70/1.

[7] R. Williams, and S. Mehta, "Leveraging Student Data to Enhance Service Delivery and Support for Sustainable Development Goal 4 in Higher Education," Journal of Quality in Higher Education, vol. 26, no. 2, 2020.

[8] A. Perez, and M. Santos, "Real-Time Sentiment Analysis for Proactive Student Support Services," Journal of Educational Technology & Society, vol. 24, no. 1, 2021.

[9] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.

[10] Y. Li, et al., "A Survey on Sentiment Analysis: From Traditional to Deep Learning Methods," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 8, 2020.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, 2019.

[12] A. Conneau, et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[14] J. Bautista, et al., "The Effectiveness of Multilingual BERT on Code-Switched Text in Southeast Asian Languages," Journal of Natural Language Engineering, vol. 28, no. 3, 2022.

[15] J. Bergstra and Y. Bengio, "Random Search for Hyper-parameter Optimization," Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.

[17] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, 2011.