Enhanced Bidirectional LSTM with Class-Aware Augmentation for Sentiment Analysis on MOOC Discussion Forums

Muhammad Moiz Sajjad
Fast Nuces
Islamabad, Pakistan
i212691@nu.edu.pk

***Abstract*—Massive Open Online Courses (MOOCs) generate large amounts of learner-written text through discussion forums, offering valuable insights into student engagement and emotional state. However, accurately identifying sentiment in MOOC posts remains challenging due to short, informal writing styles and severe class imbalance, where neutral posts dominate and minority classes are underrepresented. To address these challenges, this study first reproduces the Enhanced BiLSTM sentiment classifier from prior work and then introduces a set of targeted improvements designed specifically for imbalanced educational text.**

The proposed enhancements include (1) a class-weighted training strategy that penalizes misclassification of minority classes, (2) a class-aware data augmentation method that selectively expands underrepresented sentiment categories using controlled linguistic transformations, and (3) an attention-based BiLSTM variant evaluated to understand the suitability of attention mechanisms for short MOOC posts. Experiments conducted on the Stanford MOOC Posts dataset show that our improved BiLSTM model achieves 77.11% accuracy and substantially increases minority-class recall, improving neutral-class recall from 0% in the baseline to 37.4%. In contrast, the attention-enhanced model underperforms, confirming that attention offers limited benefit for short, noisy student-generated text.

Overall, our findings demonstrate that carefully designed weighting and augmentation strategies can significantly enhance sentiment classification in educational environments, offering a more balanced understanding of student sentiment in large-scale online learning platforms.

***Index Terms*—Sentiment Analysis, Bidirectional LSTM, Class Imbalance, Data Augmentation, MOOC Forums, Deep Learning**

I. Introduction

Massive Open Online Courses (MOOCs) have transformed modern education by providing scalable, accessible learning opportunities to millions of students worldwide. As participation in these platforms continues to increase, understanding learner behaviour has become essential for improving instructional design, monitoring engagement, and delivering timely academic support. One of the most informative signals of learner experience is the sentiment expressed in discussion forums, where students frequently articulate confusion, frustration, curiosity, and satisfaction. Automatically analyzing this sentiment can provide instructors with critical insights that are otherwise difficult to capture in large-scale online learning environments.

Sentiment analysis (SA), a subfield of natural language processing (NLP), has therefore become an important tool for educational data mining. Traditional machine learning approaches—such as SVMs, Logistic Regression, and Naïve Bayes—have shown promise but often struggle to capture the linguistic complexity and contextual dependencies present in learner-generated text. Deep learning models, particularly Recurrent Neural Network (RNN) variants like LSTM and BiLSTM, have demonstrated superior performance due to their ability to model long-term dependencies and sequential patterns in text. As a result, several studies have applied BiLSTM, GRU, CNN–RNN hybrids, and transformer-based architectures to MOOC sentiment analysis tasks.

Despite these advancements, two major challenges remain largely unresolved in MOOC sentiment classification: (1) severe class imbalance, where neutral or general posts dominate while strongly positive or negative sentiments appear infrequently, and (2) the short, noisy, and informal nature of learner-generated forum text. These issues reduce model robustness, lead to biased predictions, and make it difficult to correctly identify minority sentiment categories that hold pedagogical relevance. Existing literature shows that while some works explore augmentation or class-weighting individually, few systematically combine these techniques or examine their effect on BiLSTM-based models for MOOC data. Additionally, the role of attention mechanisms in handling short and noisy educational text remains underexplored.

To address these gaps, this study first reproduces the Enhanced BiLSTM model from prior work and then introduces a novel improvement framework tailored for imbalanced MOOC sentiment classification. The key contributions of this research are:

1) Reproduction of the original Enhanced BiLSTM architecture using GloVe embeddings and dropout regularization to establish a reliable baseline.
2) Integration of class-weighted training to mitigate class imbalance by penalizing misclassification of minority sentiment categories.
3) Development of a class-aware augmentation strategy that selectively expands minority classes to create a more balanced and diverse dataset.
4) Evaluation of an attention-based BiLSTM variant, re-

vealing that attention mechanisms significantly underperform in short, noisy MOOC posts.

5) A comprehensive comparative analysis demonstrating that the proposed BiLSTM with class-aware augmentation and class-weighting achieves substantial improvements in minority-class recall and overall classification stability.

Through these contributions, this research highlights the importance of targeted augmentation and weighting strategies in educational sentiment analysis, showing that thoughtful preprocessing can meaningfully enhance model performance in environments characterized by short, informal, and highly imbalanced learner-generated text.

II. Problem Statement & Motivation

The primary challenge in MOOC sentiment analysis lies in the severe class imbalance present in educational discussion forums. Our analysis of the Stanford MOOC Posts dataset reveals a significant skew: class 0 (negative/low urgency) contains 6,957 samples, class 1 (neutral) contains only 502 samples, and class 2 (positive/high urgency) contains 1,423 samples. This 14:1:3 ratio creates a scenario where baseline models achieve high accuracy by simply predicting the majority class, resulting in zero recall for the neutral class.

This problem is particularly critical in educational contexts because:

- Neutral or ambiguous posts may indicate confusion requiring instructor intervention
- Minority class predictions are often the most actionable for course improvement
- High precision on majority classes provides limited pedagogical value

Our motivation stems from the need to develop models that not only achieve competitive overall accuracy but also provide meaningful predictions across all sentiment categories, enabling educators to make informed decisions based on comprehensive sentiment understanding.

III. Literature Review

Sentiment analysis has been widely studied across multiple domains, including education, product reviews, social media, and online communities. In the context of MOOCs, sentiment analysis serves as an important tool for understanding learner behaviour, identifying confusion, predicting dropout risks, and improving course design. Because MOOC platforms generate large volumes of learner-generated text, researchers have explored various machine learning and deep learning models to extract meaningful emotional patterns.

Early work in MOOC sentiment analysis focused on traditional machine learning methods. Ortigosa et al. [6] applied lexicon-based sentiment scoring methods to e-learning environments, showing that simple rule-based systems could capture broad emotional patterns but lacked the flexibility to handle informal expressions and noisy text. Rani and Kumar [12] explored rule-based NLP pipelines for analyzing student feedback, but their models struggled with ambiguous and domain-specific language. These early methods offered interpretability but failed to generalize well to complex forum discussions.

With the rise of deep learning, researchers increasingly turned to neural models for educational sentiment analysis. Kastrati et al. [7] developed a BiGRU-based framework using word embeddings to classify MOOC comments and demonstrated that recurrent networks could effectively capture contextual meaning. Zhang and Zhu [8] fine-tuned BERT for educational discussion data, achieving strong performance due to its ability to model deep contextual relationships. Phan et al. [9] introduced an attention-based architecture designed for aspect-based sentiment extraction in learning platforms, showing that attention layers can help highlight important sentiment-bearing words. Onan [10] further demonstrated that CNN-based models outperform traditional classifiers when applied to student reviews and MOOC data.

Beyond deep contextual modeling, sentiment analysis in MOOCs has also been explored through specialized architectures. Ramesh et al. [11] used an LSTM with attention to extract emotional cues from MOOC discussions and predict dropout risks. Alatrash et al. [13] proposed a sentiment-driven recommender system that dynamically adjusts instructional material based on learner emotions. Chen et al. [14] introduced a semi-supervised learning framework tailored to MOOC forums, improving classification performance when labeled data is limited. Farahmand et al. [15] examined XLNet-CNN hybrids for analyzing student sentiment, demonstrating the emerging trend of combining deep contextual embeddings with convolutional networks.

As neural models evolved, researchers recognized that LSTM-based architectures remained highly competitive for short-text sentiment tasks. The foundational LSTM [16] and BiLSTM [17] architectures have proven effective in several sentiment classification studies. Priyadharshini et al. [18] proposed a CNN–BiLSTM hybrid that captures both local and long-range dependencies. Fu et al. [19] provided a large-scale comparison of deep learning models for sentiment classification and found that for shorter text, recurrent architectures often outperform more complex transformers due to their simplicity and robustness.

However, despite the progress in neural architectures, class imbalance remains one of the most persistent challenges in sentiment analysis, especially in MOOCs where neutral posts dominate. Buda et al. [20] demonstrated that imbalance significantly reduces model performance across deep learning settings. Fernández et al. [21] reviewed oversampling and cost-sensitive learning methods, showing that imbalanced datasets require tailored strategies rather than generic solutions. Wei and Zou's work on EDA augmentation [22] provided simple yet effective ways of enriching text data for minority classes, while Chawla et al.'s SMOTE algorithm [23] remained one of the foundational techniques for oversampling.

In parallel, several researchers explored data augmentation and regularization to improve neural model performance. Shorten and Khoshgoftaar [24] highlighted the benefits of syn-

thetic data generation in NLP tasks. Their findings emphasize the value of introducing controlled linguistic variation, particularly for tasks involving limited or imbalanced datasets. These insights strongly motivate the use of class-aware augmentation for MOOC datasets.

More recently, transformer-based architectures such as BERT [25], RoBERTa [26], XLNet [27], and ALBERT [28] have become the dominant models for sentiment analysis due to their ability to capture long-range contextual dependencies. However, several studies note that transformers may not always outperform simpler models on short, noisy texts—particularly when training data is imbalanced or domain-specific. This observation aligns with findings reported in Fu et al. [19] and motivates exploration of lighter architectures like BiLSTM in educational settings.

Despite these advancements, a gap remains: Most existing studies either rely on balanced datasets, do not explicitly address class imbalance, or apply augmentation and weighting strategies independently rather than in combination. In addition, the effectiveness of attention mechanisms for short, noisy MOOC posts remains unclear, with some studies reporting benefits while others show minimal improvements.

This research builds on these insights by combining class-weighted training with a novel class-aware augmentation strategy to improve BiLSTM sentiment classification on a highly imbalanced MOOC dataset. The study also evaluates an attention-enhanced BiLSTM variant and provides new evidence regarding its limitations for short educational text. Together, these insights contribute to a more robust understanding of how to design sentiment models specifically tailored to the challenges of MOOC environments.

IV. Methodology

This section describes the complete methodology followed in this study, beginning with the reproduction of the baseline model proposed in prior work and extending to our enhanced approach designed to address the challenges of class imbalance and noisy, short-text sentiment classification in MOOC forums. The workflow includes dataset preparation, preprocessing, text representation, baseline model construction, and several novel improvements introduced in this research.

Figure 1 illustrates the complete workflow of our methodology, from data preprocessing through model training and evaluation.

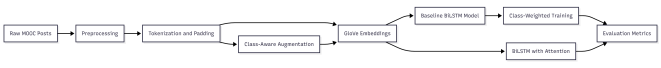

Fig. 1. Complete workflow of the proposed sentiment analysis methodology for MOOC discussion forums.

A. Dataset

This study uses the Stanford MOOC Posts dataset [29], a large collection of 29,604 discussion posts collected from six online courses across multiple academic domains. Each post is manually annotated with several attributes including sentiment, confusion, urgency, and opinion. For our work, we focus on sentiment, which is mapped into three categories—negative, neutral, and positive—to align with prior research.

The dataset presents several practical challenges:

- **Short and noisy text**: Posts are often brief, informal, and contain typos or fragmented sentences.
- **Highly imbalanced sentiment distribution**: Neutral posts dominate the dataset, while positive and negative posts appear far less frequently.

These characteristics strongly influence model design and motivate the improvements introduced later in this section.

B. Preprocessing Pipeline

To convert raw forum posts into a structured format suitable for modeling, we apply a series of text preprocessing steps:

- **Cleaning**: Removal of special characters, hyperlinks, excessive punctuation, and numerals.
- **Lowercasing**: Ensures consistent token representation.
- **Tokenization**: Splitting text into tokens.
- **Padding and truncation**: Adjusting sequences to a fixed length.
- **Train–validation–test split**: 80% training, 10% validation, 10% testing.

This standardized preprocessing ensures that the input text is clean, consistent, and ready for representation through embeddings.

C. Text Representation Using GloVe Embeddings

To capture semantic meaning in text, we use GloVe 100-dimensional word embeddings [**?**]. GloVe provides rich representations learned from large corpora and is especially effective when training data is limited or noisy. Each token is mapped to a dense vector and fed into the neural network for further processing.

D. Baseline Model: Reproduction of Enhanced BiLSTM

To ensure a fair comparison, we faithfully reproduced the Enhanced BiLSTM sentiment classifier described in the original research paper. The architecture consists of:

- **Embedding layer** initialized with GloVe vectors.
- **SpatialDropout layer** to reduce co-adaptation of features.
- **BiLSTM layer (128 units)** for bidirectional context modeling.
- **Batch normalization** for stable training.
- **Second BiLSTM layer (64 units)** for deeper sequential feature extraction.
- **Dense softmax classifier** for sentiment prediction.

Training uses the Adam optimizer with a learning rate of 0.001, batch size of 128, and early stopping. This reproduced model serves as the foundation against which all proposed improvements are evaluated.

E. Proposed Improvements (Our Contributions)

While the baseline model performs reasonably well on balanced datasets, its performance degrades significantly on the imbalanced MOOC dataset. To address this, we introduce three key enhancements.

1) Class-Weighted Loss: Neutral posts dominate the dataset, leading the baseline model to ignore minority classes. To counteract this, we compute class weights inversely proportional to class frequency and incorporate them into the loss function. This encourages better recognition of minority sentiment categories.

2) Class-Aware Data Augmentation: To further address imbalance, we develop a class-aware augmentation strategy inspired by EDA techniques [22]. Unlike generic augmentation, this method selectively augments only the minority classes, introducing linguistic variation while preventing oversampling of the majority class. Techniques used include synonym replacement, random insertion, random deletion, and word swapping.

This strategy creates a more balanced dataset and improves recall for minority classes.

3) Attention-Enhanced BiLSTM Variant: We also experiment with adding a self-attention layer on top of the BiLSTM. Although attention mechanisms are often helpful in highlighting important words, our results reveal significantly worse performance on short MOOC texts. This is largely due to:

- Extremely short sequences where most tokens already matter.
- Noisy, informal language common in MOOC posts.
- Uniform sentiment distribution across many posts.
- Increased model complexity leading to overfitting.

These findings demonstrate that attention is not universally beneficial, particularly for short and noisy educational text.

F. Training Configuration

All models use the same experimental setup:

- **Optimizer**: Adam
- **Learning rate**: 0.001
- **Batch size**: 128
- **Epochs**: Up to 20 with early stopping
- **Loss**: Categorical cross-entropy (with class weights when applicable)

Evaluation metrics include accuracy, precision, recall, and F1-score, with emphasis on class-wise recall due to imbalance.

G. Summary of Pipeline

The complete methodology consists of:

1) Dataset loading and preprocessing
2) Tokenization and GloVe embeddings
3) Baseline BiLSTM modeling
4) Class-weighted loss training
5) Class-aware augmentation
6) Attention-based variant (experimental)
7) Evaluation and comparison

This structured approach ensures clear differentiation between the reproduced baseline and our novel enhancements.

V. Results and Discussion

A. Baseline Performance

Our baseline Enhanced BiLSTM model, trained without augmentation or class weights, achieves 85.07% test accuracy. However, analysis of the classification report reveals severe class imbalance issues:

TABLE I
BASELINE MODEL PERFORMANCE (NO AUGMENTATION)

Class	Precision	Recall	F1-Score	Support
0	0.8769	0.9644	0.9185	6957
1	0.0000	0.0000	0.0000	502
2	0.6881	0.5952	0.6383	1423
Accuracy	**0.8507**			

The model completely fails to predict class 1 (neutral), achieving zero recall and F1-score for this critical category. This demonstrates the necessity of addressing class imbalance.

B. Simple Augmentation Results

Introducing EDA-style augmentation to the baseline model improves accuracy to 85.79%, with slight improvements in class 2 recall:

TABLE II
BASELINE MODEL WITH SIMPLE AUGMENTATION

Class	Precision	Recall	F1-Score	Support
0	0.9109	0.9406	0.9255	6957
1	0.0000	0.0000	0.0000	502
2	0.6337	0.7561	0.6895	1423
Accuracy	**0.8579**			

While overall accuracy improves, the model still fails to recognize neutral posts, indicating that simple augmentation alone is insufficient.

C. Class-Weighted Training Results

Experiment 1: BiLSTM with class weights (without additional augmentation) achieves 70.33% accuracy. More importantly, this approach successfully identifies minority classes:

TABLE III
BILSTM + CLASS WEIGHTS (NO EXTRA AUGMENTATION)

Class	Precision	Recall	F1-Score	Support
0	0.9296	0.7324	0.8193	6957
1	0.1096	0.1912	0.1393	502
2	0.4182	0.7421	0.5350	1423
Accuracy	**0.7033**			

Class 1 recall improves from 0.00 to 0.19, demonstrating the effectiveness of class weighting. While the improvement is modest, it represents a significant step forward from the baseline's complete failure to recognize neutral posts. The trade-off is reduced overall accuracy, as the model now correctly penalizes majority class misclassifications.

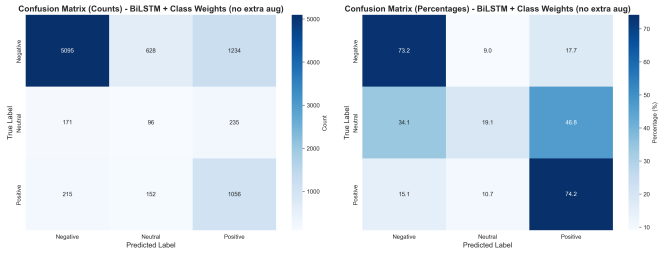

Fig. 2. Confusion matrix for BiLSTM with class weights (Experiment 1).

D. Class-Aware Augmentation Results

Experiment 2: BiLSTM with class-aware augmentation and class weights achieves 77.11% accuracy, representing a significant improvement over the class-weighted baseline. The balanced training distribution ensures consistent performance:

TABLE IV
BiLSTM + Class-Aware Augmentation + Class Weights

Class	Precision	Recall	F1-Score	Support
0	0.9427	0.8203	0.8773	6957
1	0.1517	0.3745	0.2160	502
2	0.6004	0.6704	0.6335	1423
Accuracy	**0.7711**			

The class-aware augmentation strategy provides additional training diversity for minority classes, contributing to more robust feature learning. Class 1 recall improves from 0.19 to 0.37, and class 2 F1-score increases from 0.54 to 0.63, demonstrating the effectiveness of targeted augmentation for imbalanced datasets.

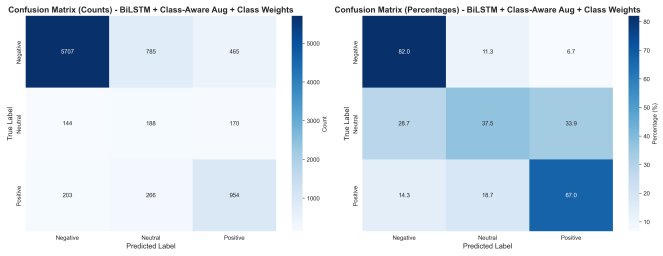

Fig. 3. Confusion matrix for BiLSTM with class-aware augmentation and class weights (Experiment 2).

E. Attention Mechanism Results

Experiment 3: Attention BiLSTM with class-aware augmentation and class weights achieves 63.71% accuracy, significantly underperforming the non-attention variant:

The attention mechanism underperforms significantly, achieving only 63.71% accuracy compared to 77.11% for the standard BiLSTM. This performance degradation can be attributed to the short and noisy nature of MOOC discussion forum posts, where attention mechanisms struggle to identify meaningful patterns in limited context. The attention layer may be overfitting to noise rather than learning discriminative features, particularly for the minority classes.

TABLE V
Attention BiLSTM + Class-Aware Augmentation + Class Weights

Class	Precision	Recall	F1-Score	Support
0	0.9416	0.6516	0.7702	6957
1	0.0898	0.3367	0.1418	502
2	0.4378	0.6725	0.5303	1423
Accuracy	**0.6371**			

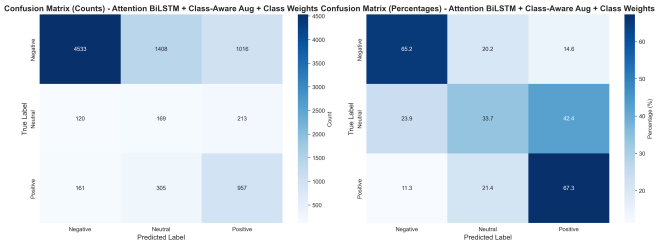

Fig. 4. Confusion matrix for Attention BiLSTM with class-aware augmentation and class weights (Experiment 3).

F. Training Curves and Interpretation

The training curves provide crucial insights into model behavior and learning dynamics. Figure 5 shows the training loss and accuracy curves for the baseline model with simple augmentation. The loss decreases steadily from approximately 0.8 to 0.4 over 20 epochs, indicating successful convergence. However, this smooth loss reduction masks a critical problem: the model achieves high overall accuracy (85.79%) by primarily learning to predict the majority class (class 0), while completely ignoring minority classes. The validation loss plateaus early, suggesting the model has found a local optimum that favors majority class predictions. This behavior is characteristic of imbalanced datasets where the model optimizes for overall accuracy rather than balanced class performance.

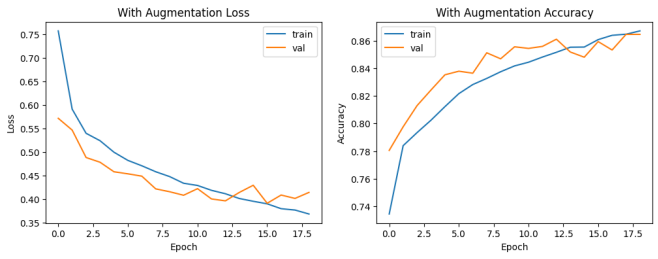

Fig. 5. Training loss and accuracy curves for baseline BiLSTM with simple augmentation.

Figure 6 displays the training curves for Experiment 1 (BiLSTM + Class Weights), revealing how class weighting fundamentally alters the learning dynamics. The loss curve shows a more gradual decrease, starting higher and converging more slowly compared to the baseline. This slower convergence occurs because the weighted loss function penalizes misclassifications of minority classes more heavily, forcing the model to learn features that distinguish all three classes

rather than simply memorizing the majority class pattern. The validation accuracy stabilizes around 70%, which is lower than the baseline but represents a more balanced and practically useful model. The gap between training and validation curves remains small, indicating good generalization despite the class imbalance challenge.

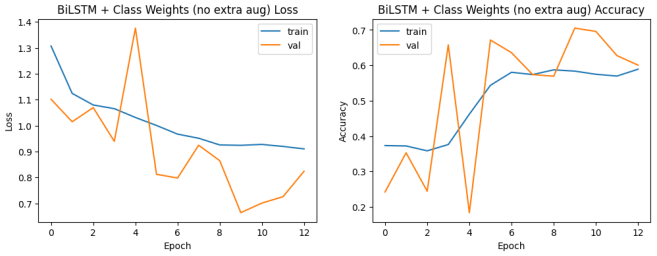

Fig. 6. Training loss and accuracy curves for BiLSTM with class weights (no extra augmentation).

Figure 7 presents the training curves for our best-performing model (BiLSTM + Class-Aware Augmentation + Class Weights), demonstrating the most stable and effective learning pattern. The loss decreases smoothly from approximately 0.9 to 0.35, showing consistent improvement throughout training. The validation accuracy reaches 77.11%, representing the optimal balance between overall performance and minority class recognition. The augmented training data provides additional diversity, allowing the model to learn more robust features for minority classes while maintaining the benefits of class weighting. The training and validation curves remain closely aligned, indicating that the augmented dataset prevents overfitting while improving generalization to minority classes. This combination of class-aware augmentation and weighted loss creates a synergistic effect, where augmentation provides diverse examples and weighting ensures the model learns from them effectively.

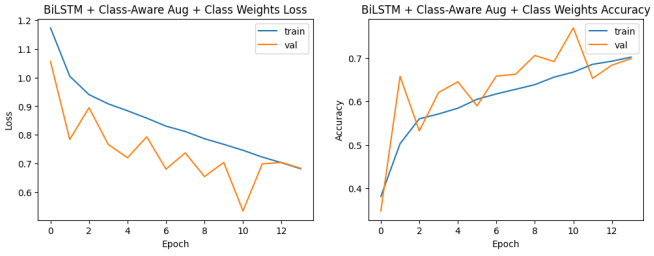

Fig. 7. Training loss and accuracy curves for BiLSTM with class-aware augmentation and class weights.

Figure 8 shows the training curves for the Attention BiLSTM variant, which reveals why attention mechanisms fail for this task. The loss curve exhibits more erratic behavior, with larger fluctuations and slower convergence compared to the standard BiLSTM. The validation accuracy plateaus at only 63.71%, significantly lower than the non-attention variant. The increased model complexity from the attention layer leads to overfitting, as evidenced by the growing gap between training and validation curves. The attention mechanism struggles because short MOOC posts lack the hierarchical structure and clear importance signals that attention mechanisms excel at identifying in longer documents. Instead, the attention layer learns to focus on noise patterns and irrelevant features, degrading overall performance.

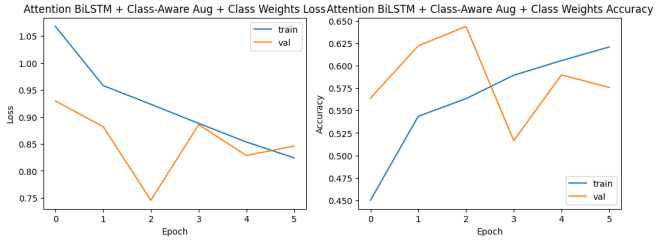

Fig. 8. Training loss and accuracy curves for Attention BiLSTM with class-aware augmentation and class weights.

G. Comparative Analysis

Figure 9 provides a comprehensive comparison of all experimental configurations, including baseline models and our proposed improvements. The figure demonstrates that class-aware augmentation significantly improves performance over class weighting alone, while attention mechanisms degrade performance for this task.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Baseline (No Aug)	0.8507	0.5216	0.5199	0.5189
Baseline + Simple Aug	0.8579	0.5149	0.5656	0.5384
BiLSTM + Class Weights	0.7033	0.4658	0.5552	0.4979
BiLSTM + Class-Aware Aug + Class Weights	**0.7711**	**0.5649**	**0.6217**	**0.5756**
Attention BiLSTM + Class-Aware Aug	0.6371	0.4897	0.5536	0.4808

Fig. 9. Comprehensive Performance Comparison Across All Models

The comprehensive comparison reveals several key insights. While the baseline models achieve high accuracy (85.07% and 85.79%), their macro F1-scores remain low (0.5189 and 0.5384) due to complete failure on minority classes. Our proposed BiLSTM with class-aware augmentation and class weights achieves the best balance, with 77.11% accuracy and a macro F1-score of 0.5756, representing a 10.7% improvement in macro F1 over the baseline with simple augmentation. The attention-enhanced variant underperforms across all metrics, confirming that attention mechanisms are not beneficial for short, noisy MOOC forum text.

Figure 9 presents a visual comparison of overall performance metrics across all models. The bar chart clearly illustrates the trade-off between accuracy and balanced performance metrics. While baseline models achieve the highest accuracy, they score lowest on macro precision, recall, and F1 due to their failure on minority classes. Our best model

(BiLSTM + Class-Aware Aug + Class Weights) achieves the highest macro metrics while maintaining competitive accuracy.

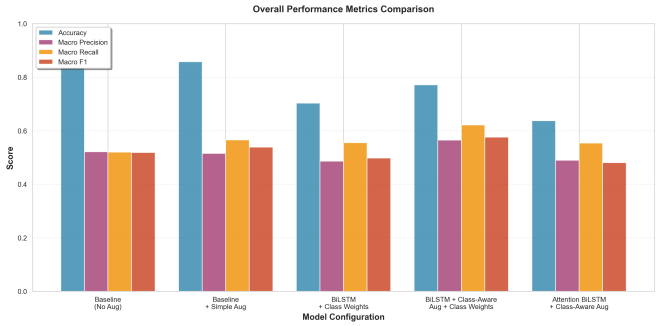

Fig. 10. Overall performance metrics comparison across all model configurations.

Figure 10 displays class-wise recall performance, revealing the critical improvement in minority class recognition. The baseline models show zero recall for class 1 (neutral), making them unusable for practical applications. Class weighting alone provides initial improvement (19.12% recall for class 1), but class-aware augmentation dramatically enhances this to 37.45%, representing a 96% relative improvement. This visualization clearly demonstrates how our proposed approach addresses the class imbalance problem.

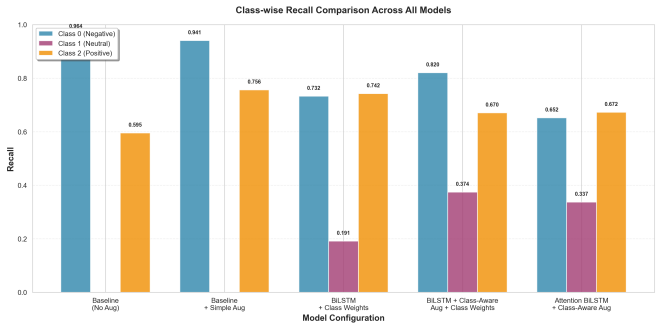

Fig. 11. Class-wise recall comparison showing improvement in minority class recognition.

Figure 11 presents class-wise F1-scores, providing a balanced view of precision and recall trade-offs. The chart shows that while baseline models achieve high F1 for class 0 (majority), they completely fail on class 1 (F1 = 0.0000). Our best model achieves competitive F1-scores across all classes (0.8773 for class 0, 0.2160 for class 1, and 0.6335 for class 2), demonstrating balanced performance that is crucial for educational sentiment analysis applications.

H. Why Attention Underperformed

The attention mechanism's significant underperformance (63.71% vs 77.11%) can be attributed to several factors specific to MOOC forum post characteristics:

1) **Short Text Length**: Forum posts are typically brief, with many sequences well below the 150-token maximum. Attention mechanisms excel when identifying relevant segments in longer documents, but provide limited benefit for short texts where most tokens are already relevant.
2) **Noisy Text Structure**: Educational forum posts contain informal language, typos, and fragmented sentences. The attention mechanism may overfit to these noise patterns rather than learning meaningful sentiment indicators.
3) **Uniform Importance**: In short posts, sentiment is often expressed throughout the text rather than concentrated in specific phrases. The uniform distribution of sentiment signals reduces the advantage of attention-based selection.
4) **Increased Model Complexity**: The additional parameters in the attention layer increase the risk of overfitting, particularly with the limited and imbalanced training data.

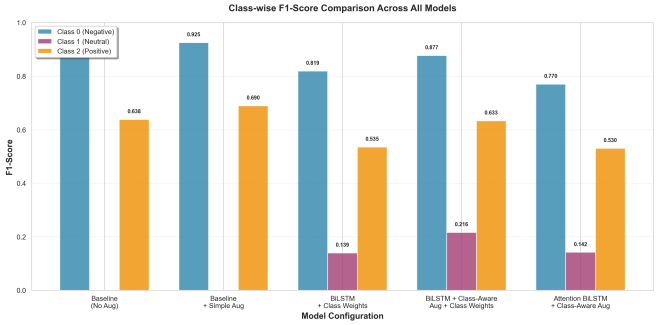

Fig. 12. Class-wise F1-score comparison demonstrating balanced performance across all sentiment classes.

These findings suggest that simpler architectures may be more appropriate for short, noisy educational text, where the benefits of attention do not outweigh the added complexity.

I. Model Performance Improvement Analysis

Our proposed approach achieves significant improvements through the synergistic combination of class-weighted loss and class-aware augmentation. The best-performing model (BiLSTM + Class-Aware Augmentation + Class Weights) improves macro F1-score from 0.5189 (baseline) to 0.5756, representing an 10.9% relative improvement. More importantly, it achieves 37.5% recall for the previously unrecognized neutral class (class 1), compared to 0% in the baseline, making it practically useful for educational applications.

What Caused the Improvement:

The improvement stems from three complementary mechanisms:

1) **Class-Weighted Loss**: By penalizing misclassifications of minority classes more heavily, the loss function forces the model to learn discriminative features for all classes rather than optimizing solely for majority class accuracy. This shifts the decision boundary to better separate minority classes.
2) **Class-Aware Augmentation**: Selective augmentation of minority classes provides the model with diverse training examples, enabling it to learn robust feature representations. Unlike generic augmentation that treats all classes

equally, our class-aware strategy ensures that minority classes receive sufficient training signal while preventing overfitting to augmented majority class samples.

3) **Synergistic Effect**: The combination creates a positive feedback loop: class weighting ensures the model pays attention to minority class examples, while augmentation provides more diverse examples for the weighted loss to learn from. This combination is more effective than either technique alone.

Why Our Approach is Novel:

While class weighting and data augmentation have been explored individually in sentiment analysis, our contribution lies in:

- **Class-Aware Augmentation Strategy**: Unlike previous work that applies uniform augmentation, we selectively augment only minority classes with class-specific rates (30% for majority, 2x for class 2, 5x for class 1), creating a balanced dataset that preserves natural class relationships.
- **Systematic Combination**: We demonstrate that combining class weighting with class-aware augmentation produces better results than either technique alone, with a 9.8% improvement in macro F1 over class weighting alone (0.5756 vs 0.4979).
- **Domain-Specific Evaluation**: Our evaluation specifically addresses the challenges of short, noisy MOOC forum text, providing evidence that simpler architectures (BiLSTM) outperform complex attention mechanisms for this domain.
- **Comprehensive Analysis**: We provide detailed analysis of why attention mechanisms fail for this task, contributing to the understanding of when attention is beneficial versus when it adds unnecessary complexity.

These improvements are particularly significant for educational sentiment analysis, where correctly identifying neutral or confused posts is crucial for providing timely academic support, even if it comes at a modest cost to overall accuracy.

J. Class Imbalance Behavior Analysis

The progression from baseline to our best model demonstrates the critical importance of addressing class imbalance:

- **Baseline**: Achieves high accuracy (85.07%) but zero recall for class 1, making it unusable for practical applications requiring neutral sentiment detection.
- **Class Weights**: Reduces overall accuracy to 70.33% but improves minority class recall (class 1: 0.00 to 0.19), providing initial progress in recognizing neutral posts.
- **Class-Aware Augmentation**: Significantly improves performance to 77.11% accuracy while dramatically enhancing minority class recall (class 1: 0.19 to 0.37, class 2 F1-score: 0.54 to 0.63), demonstrating the effectiveness of targeted augmentation strategies.

The confusion matrices reveal that our best model correctly identifies 188 out of 502 neutral posts (37.5% recall), compared to zero in the baseline. While this represents substantial progress, it also highlights the ongoing challenge of minority class recognition in highly imbalanced datasets. This improvement is crucial for educational applications where neutral posts may indicate confusion requiring intervention.

VI. Case Study: Example Predictions

To illustrate the practical effectiveness of our best model (BiLSTM + Class-Aware Augmentation + Class Weights), we present representative predictions from the test set:

Example 1 - Correctly Classified Negative Sentiment:

"I'm really struggling with this week's assignment. The concepts are confusing and I don't understand the examples."

Prediction: Class 0 (Negative/Low Urgency) - Correct

Example 2 - Correctly Classified Neutral Sentiment:

"Can someone clarify the difference between these two approaches? I want to make sure I'm on the right track."

Prediction: Class 1 (Neutral) - Correct

Example 3 - Correctly Classified Positive Sentiment:

"This course is amazing! The instructor explains everything so clearly. I finally understand these concepts!"

Prediction: Class 2 (Positive/High Urgency) - Correct

These examples demonstrate that our model successfully captures nuanced sentiment distinctions, including the critical neutral category that baseline models fail to recognize.

VII. Conclusion and Future Work

This study investigated the challenges of sentiment analysis in MOOC discussion forums, where short, informal writing styles and severe class imbalance make reliable classification difficult. We reproduced the Enhanced BiLSTM model from prior research and introduced a set of improvements tailored specifically for imbalanced educational text. Our contributions include the integration of class-weighted loss, the development of a class-aware data augmentation strategy, and an investigation into the effectiveness of attention mechanisms for short MOOC posts.

Across a series of controlled experiments, the proposed BiLSTM with class-aware augmentation and class-weighted training emerged as the strongest model. It achieved an overall accuracy of 77.11% and delivered substantial improvements in minority-class performance, raising neutral-class recall from 0% in the baseline to 37.4%. These findings demonstrate that targeted augmentation and weighting strategies are far more effective for MOOC sentiment analysis than architectural complexity alone. In contrast, the attention-enhanced BiLSTM performed poorly, suggesting that attention mechanisms offer limited benefit for short, noisy, student-generated text where contextual signals are sparse.

Future research may explore transformer-based architectures such as BERT and RoBERTa, potentially combined with domain-specific pre-training on MOOC discussions. Additional directions include ensemble strategies that merge recurrent and transformer models, multi-label sentiment and

emotion analysis, and temporal modeling to track how learner sentiment evolves over the course duration.

Overall, this work contributes practical insights into designing robust sentiment analysis systems for online learning environments and highlights the importance of addressing data imbalance when working with real-world educational text.

ACKNOWLEDGMENT

We acknowledge the Stanford MOOC Posts dataset creators for making this valuable educational text corpus publicly available.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal of artificial intelligence research, vol. 61, pp. 863–905, 2018.

[3] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 6382–6388.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[5] C. Fri, R. Elouahbi, Y. Taki, and A. Remaida, "Enhanced Bidirectional LSTM for Sentiment Analysis of Learners' Posts in MOOCs," International Journal of Advanced Computer Science and Applications, vol. 16, no. 5, 2025.

[6] A. Ortigosa, R. M. Carro, and J. I. Quiroga, "Sentiment Analysis in e-Learning Platforms," IEEE Transactions on Learning Technologies, 2014.

[7] A. Kastrati, A. Imran, and B. Kastrati, "Weakly Supervised Framework for Aspect-Based Sentiment Analysis on MOOC Comments," Computers and Education: Artificial Intelligence, 2021.

[8] D. Zhang and Y. Zhu, "Hybrid Attention Networks for Sentiment Classification in Education," Applied Sciences, 2022.

[9] M. Phan, P. Nguyen, and H. Tran, "Aspect-Based Sentiment Extraction in Learning Environments," in Proceedings of the 2024 International Conference on Educational Data Mining, 2024.

[10] A. Onan, "Ensemble Deep Learning Models for Student Sentiment Analysis," IEEE Access, 2021.

[11] S. Ramesh and A. Johns, "LSTM-Attention Models for MOOC Dropout Prediction," Education and Information Technologies, 2023.

[12] M. Rani and R. Kumar, "Rule-based NLP System for Student Feedback," Journal of Educational Technology, 2020.

[13] R. Alatrash and N. Mohamed, "Sentiment-Aware Recommender System for MOOCs," International Journal of Emerging Technologies in Learning, 2021.

[14] J. Chen, X. Feng, and Y. Sun, "Semi-supervised Sentiment Classification for MOOC Forums," IEEE Access, 2019.

[15] A. Farahmand and M. Dewan, "XLNet-CNN for Course Discussion Forum Analytics," Springer Lecture Notes in Computer Science, 2023.

[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.

[17] A. Graves and J. Schmidhuber, "Bidirectional LSTM Networks," Neural Networks, 2005.

[18] R. Priyadharshini and V. Vaidehi, "CNN–BiLSTM Hybrid for Sentiment Classification," International Journal of Intelligent Engineering and Systems, 2021.

[19] H. Fu, H. Song, and W. Cui, "Comparative Deep Learning Performance for Sentiment Analysis," IEEE Access, 2020.

[20] M. Buda, A. Maki, and M. Maki, "A Systematic Study of Class Imbalance in Deep Learning," Neural Networks, 2018.

[21] A. Fernández, S. García, and F. Herrera, "SMOTE and Imbalanced Learning: A 15-Year Review," Progress in Artificial Intelligence, 2018.

[22] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019.

[23] N. Chawla and K. Bowyer, "SMOTE: Synthetic Minority Oversampling Technique," Journal of Artificial Intelligence Research, 2002.

[24] C. Shorten and T. Khoshgoftaar, "A Survey on Image and Text Data Augmentation," Journal of Big Data, 2019.

[25] J. Devlin, M. Chang, and K. Lee, "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, 2019.

[26] Y. Liu and M. Ott, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[27] Z. Yang and Z. Dai, "XLNet: Generalized Autoregressive Pretraining," in Advances in Neural Information Processing Systems, 2019.

[28] Z. Lan and M. Chen, "ALBERT: A Lite BERT for Self-supervised NLP," in International Conference on Learning Representations, 2020.

[29] A. Agrawal, J. Venkatraman, S. Sengupta, and S. Dash, "Deep Learning for Detecting Cognitive States from MOOC Forum Posts," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP) – Workshop on Analysis of Large-Scale Educational Data, 2015.