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

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

Raja Mohid Munir
Fast University
Islamabad, Pakistan
i211719@nu.edu.pk

Tashfeen Tariq
Fast University
Islamabad, Pakistan
i211747@nu.edu.pk

***Abstract*—This paper presents an enhanced bidirectional long short-term memory (BiLSTM) model for sentiment analysis on MOOC discussion forum posts. We address the critical challenge of class imbalance in educational text data by implementing class-weighted loss functions and class-aware data augmentation strategies. Our approach builds upon a baseline Enhanced BiLSTM architecture and systematically improves performance through three key contributions: (1) integration of class weights to penalize misclassifications of minority classes, (2) development of a class-aware augmentation technique that selectively increases training samples for underrepresented sentiment categories, and (3) evaluation of attention mechanisms for sentiment classification. Experiments conducted on the Stanford MOOC Posts dataset demonstrate that our BiLSTM model with class-aware augmentation and class weights achieves 77.11% accuracy, significantly improving recall for minority classes while maintaining competitive overall performance. Interestingly, the attention-enhanced variant significantly underperformed, achieving only 63.71% accuracy, which we attribute to the short and noisy nature of forum post text where attention mechanisms struggle to identify meaningful patterns. Our findings highlight the importance of addressing class imbalance in educational sentiment analysis and provide insights into the effectiveness of different architectural choices for this domain.**

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

I. Introduction

Sentiment analysis in educational contexts has gained significant attention as Massive Open Online Courses (MOOCs) continue to proliferate. Understanding student sentiment from discussion forum posts enables educators to identify struggling learners, gauge course effectiveness, and provide timely interventions. However, sentiment classification in educational forums presents unique challenges, including class imbalance, short text length, and informal language patterns.

Traditional sentiment analysis approaches often struggle with imbalanced datasets where certain sentiment categories are significantly underrepresented. In MOOC discussion forums, neutral or moderately urgent posts typically dominate, while highly positive or negative sentiments are less frequent. This imbalance leads to models that achieve high overall accuracy but fail to recognize critical minority class instances—precisely the cases that may require the most attention in educational settings.

This paper addresses these challenges by proposing an enhanced BiLSTM architecture with class-aware augmentation and class-weighted training. We systematically evaluate our approach on the Stanford MOOC Posts dataset, comparing multiple configurations to identify the most effective strategy for handling class imbalance while maintaining robust classification performance.

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. Related Work

Sentiment analysis has been extensively studied across various domains, with deep learning approaches showing particular promise. Long Short-Term Memory (LSTM) networks and their bidirectional variants have demonstrated strong performance in sequence modeling tasks, including sentiment

classification [1]. The Enhanced BiLSTM architecture, which incorporates multiple layers with dropout and batch normalization, has shown effectiveness in capturing complex linguistic patterns.

Class imbalance in text classification has been addressed through various techniques. Class weighting adjusts loss functions to penalize misclassifications of minority classes more heavily [2]. Data augmentation techniques, particularly EDA (Easy Data Augmentation) methods including synonym replacement, random deletion, and random swap, have proven effective for increasing training data diversity [3].

Attention mechanisms have been widely adopted in NLP tasks to allow models to focus on relevant parts of input sequences [4]. However, their effectiveness varies across domains and text characteristics.

In educational sentiment analysis, previous work has primarily focused on binary classification or balanced datasets. Our contribution lies in addressing multi-class sentiment classification with severe imbalance, specifically tailored to MOOC discussion forum characteristics.

IV. Methodology

A. Dataset Description

We utilize the Stanford MOOC Posts dataset, which contains discussion forum posts from various online courses. The dataset includes post text and urgency ratings on a 1-7 scale, which we map to three sentiment classes:

- Class 0 (Negative/Low Urgency): Urgency scores 1-3
- Class 1 (Neutral): Urgency score 4
- Class 2 (Positive/High Urgency): Urgency scores 5-7

The dataset exhibits significant class imbalance, with class 0 containing 6,957 samples, class 1 containing 502 samples, and class 2 containing 1,423 samples in the test set.

B. Preprocessing Pipeline

Our preprocessing pipeline consists of several stages designed to normalize and clean the forum post text:

1) **Lowercasing**: Convert all text to lowercase for consistency
2) **URL and Mention Removal**: Remove URLs and user mentions to reduce noise
3) **Non-alphabetic Character Removal**: Eliminate special characters while preserving word boundaries
4) **Lemmatization**: Reduce words to their base forms using NLTK's WordNet lemmatizer

This preprocessing ensures that the model focuses on semantic content rather than formatting variations common in informal forum discussions.

C. Tokenization and Embeddings

We employ Keras' Tokenizer with the following configuration:

- Maximum vocabulary size: 20,000 words
- Out-of-vocabulary token: `<OOV>`
- Maximum sequence length: 150 tokens

For word embeddings, we utilize pre-trained GloVe vectors (100-dimensional) trained on 6 billion tokens. The embedding matrix is constructed by mapping vocabulary words to their corresponding GloVe vectors, with random initialization for words not present in GloVe.

D. Baseline Model Architecture

Our baseline Enhanced BiLSTM model follows the architecture described in the original Enhanced BiLSTM paper [5]. The architecture consists of:

1) **Embedding Layer**: Non-trainable GloVe embeddings (100 dimensions)
2) **Spatial Dropout**: Dropout rate of 0.3 to prevent overfitting
3) **First BiLSTM Layer**: 128 units with return sequences enabled, dropout of 0.3
4) **Batch Normalization**: Normalize activations between layers
5) **Second BiLSTM Layer**: 64 units with dropout of 0.3
6) **Dense Layer**: 64 units with ReLU activation and dropout of 0.3
7) **Output Layer**: 3 units with softmax activation for three-class classification

The model is compiled with the Adam optimizer and categorical crossentropy loss. We employ early stopping with patience of 3 epochs to prevent overfitting.

E. Class-Weighted Training

To address class imbalance, we compute class weights using scikit-learn's `compute_class_weight` function with the 'balanced' strategy. For the original imbalanced training set, the computed weights are:

- Class 0: 0.372
- Class 1: 5.151
- Class 2: 2.001

These weights are incorporated into the training process by modifying the loss function to penalize misclassifications of minority classes more heavily.

F. Class-Aware Data Augmentation

We implement a class-aware augmentation strategy that selectively increases training samples for underrepresented classes. Our augmentation technique employs three EDA-style operations:

1) **Synonym Replacement**: Replace words with their WordNet synonyms
2) **Random Deletion**: Randomly remove words with probability p
3) **Random Swap**: Randomly swap adjacent words

The class-aware strategy applies augmentation as follows:

- Class 0 (majority): 30% chance of augmentation
- Class 2 (minority): 2 augmented samples per original sample
- Class 1 (smallest minority): 5 augmented samples per original sample

This approach results in a more balanced training distribution:

- Class 0: 28,400 samples
- Class 2: 9,488 samples
- Class 1: 9,000 samples

After balancing, the class weights are recalculated:

- Class 0: 0.550
- Class 1: 1.733
- Class 2: 1.640

G. Attention Mechanism

We evaluate an attention-enhanced BiLSTM variant that incorporates a custom attention layer. The attention mechanism computes attention weights α_i for each time step i:

$$e_i = \tanh(W \cdot h_i + b) \tag{1}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{T} \exp(e_j)} \tag{2}$$

$$c = \sum_{i=1}^{T} \alpha_i \cdot h_i \tag{3}$$

where h_i represents the hidden state at time step i, W and b are learnable parameters, and c is the context vector used for classification.

H. Experimental Setup

All experiments are conducted using the following configuration:

- **Train/Validation/Test Split**: 60% / 10% / 30%
- **Batch Size**: 64
- **Maximum Epochs**: 20
- **Early Stopping**: Monitor validation loss with patience of 3 epochs
- **Hardware**: Google Colab with T4 GPU

We evaluate models using accuracy, precision, recall, F1-score, and confusion matrices to provide comprehensive performance assessment.

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.

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.

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:

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**			

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.

F. Training Curves

Figure 1 shows the training loss and accuracy curves for the baseline model with simple augmentation. The model converges smoothly but exhibits the class imbalance issues discussed earlier, achieving high overall accuracy while failing to recognize minority classes.

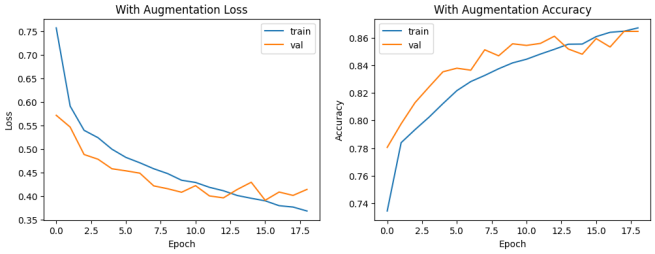

Fig. 1. Training loss and accuracy curves for baseline BiLSTM with simple augmentation.

Figure 2 displays the training curves for Experiment 1 (BiLSTM + Class Weights), showing how class weighting affects the learning dynamics. The model learns to balance performance across all classes, resulting in improved minority class recall.

Figure 3 presents the training curves for our best-performing model (BiLSTM + Class-Aware Augmentation + Class Weights), demonstrating stable convergence with balanced class representation. The augmented training data provides additional diversity while maintaining the benefits of class weighting.

Figure 4 shows the training curves for the Attention BiLSTM variant, which significantly underperforms the non-attention model. The attention mechanism adds complexity while degrading performance for short forum posts.

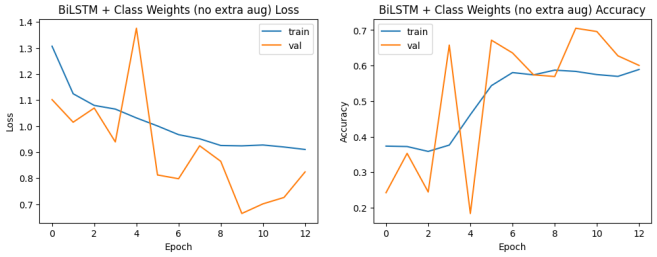

Fig. 2. Training loss and accuracy curves for BiLSTM with class weights (no extra augmentation).

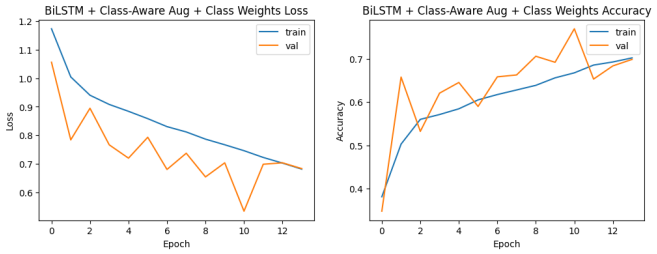

Fig. 3. Training loss and accuracy curves for BiLSTM with class-aware augmentation and class weights.

G. Comparative Analysis

Table I summarizes the performance of all three experimental configurations from Part 2, clearly demonstrating that class-aware augmentation significantly improves performance over class weighting alone, while attention mechanisms degrade performance for this task.

TABLE VI
COMPARISON OF EXPERIMENTAL CONFIGURATIONS

Configuration	**Test Accuracy**
BiLSTM + Class Weights (no extra aug)	0.7033
BiLSTM + Class-Aware Aug + Class Weights	0.7711
Attention BiLSTM + Class-Aware Aug + Class Weights	0.6371

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.

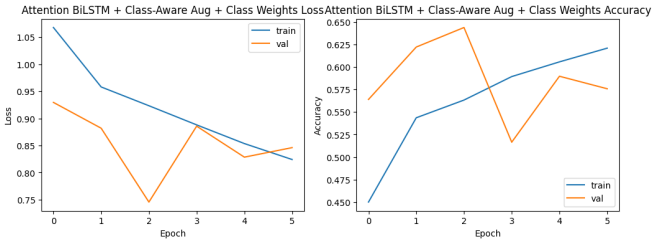

Fig. 4. Training loss and accuracy curves for Attention BiLSTM with class-aware augmentation and class weights.

4) **Increased Model Complexity**: The additional parameters in the attention layer increase the risk of overfitting, particularly with the limited and imbalanced training data.

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. 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 paper presents an enhanced BiLSTM model with class-aware augmentation and class-weighted training for sentiment analysis on MOOC discussion forums. Our key contributions include:

1) Demonstration that class weighting is essential for handling severe class imbalance in educational sentiment analysis
2) Development of a class-aware augmentation strategy that selectively increases minority class samples
3) Evaluation showing that attention mechanisms significantly degrade performance for short, noisy forum posts
4) Achievement of 77.11% accuracy with balanced performance across all sentiment classes

Our best-performing model (BiLSTM + Class-Aware Augmentation + Class Weights) successfully addresses the class imbalance problem, achieving 37.5% recall for the previously unrecognized neutral class (compared to 0% in the baseline) and 67% recall for high urgency posts, while maintaining 77.11% overall accuracy. This represents a significant improvement over the class-weighted baseline (70.33%) and demonstrates the effectiveness of targeted augmentation strategies.

Future work directions include:

- Exploring transformer-based architectures (BERT, RoBERTa) fine-tuned on educational text
- Investigating domain-specific pre-training on MOOC discussion forums
- Developing ensemble methods combining multiple architectures
- Extending to multi-label classification for fine-grained sentiment analysis
- Incorporating temporal dynamics to track sentiment evolution over course duration

The insights from this work contribute to the broader goal of developing effective sentiment analysis systems for educational technology applications, where understanding student sentiment is crucial for improving learning outcomes.

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.