International Society of the Learning Sciences

Balancing the Imbalance: Enhancing MOOC Discussion Forum Urgency Prediction with LLM-Generated Data Augmentation

Hongming Li, Anthony F. Botelho
hli3@ufl.edu, abotelho@coe.ufl.edu
University of Florida

Abstract: Discussion forums in Massive Open Online Courses (MOOCs) often contain urgent posts requiring timely instructor intervention, but identifying these posts is challenging due to severe class imbalance in urgency levels. This study presents a novel approach combining Large Language Models (LLMs) and traditional machine learning to address this challenge. We develop a two-stage pipeline that first uses LLMs to generate synthetic forum posts, creating a balanced dataset across seven urgency levels, then evaluates five different machine learning models on both original and balanced datasets. Our results demonstrate significant improvements across all models, with the Support Vector Regression achieving the highest performance (R^2 improvement from 0.298 to 0.857). Through comprehensive evaluation using multiple metrics and cross-validation, we show that our approach not only improves model performance but also enhances prediction stability across different urgency levels. This work contributes to both educational data mining methodology and practical MOOC forum management.

Introduction

Massive Open Online Courses (MOOCs) have transformed educational accessibility, attracting millions of learners worldwide. However, the scale that makes MOOCs revolutionary also creates significant challenges in providing timely support to learners (Almatrafi et al., 2018). Discussion forums, while crucial for learner engagement and support, often become overwhelming for instructors due to the high volume of posts and varying levels of urgency (Almatrafi et al., 2018; Guo et al., 2019).

A particularly critical challenge is identifying posts that require urgent instructor intervention. These posts, if left unaddressed, can significantly impact learner progression and potentially contribute to the notably high dropout rates in MOOCs (El-Rashidy et al., 2024). However, automating the identification of urgent posts presents a fundamental data science challenge: severe class imbalance. In typical MOOC forums, highly urgent posts are rare but critically important, creating a "needle in a haystack" problem that compromises traditional machine learning approaches (Almatrafi et al., 2018; Alrajhi et al., 2024).

Recent advances in Large Language Models (LLMs) offer promising new directions for addressing data imbalance through synthetic data generation. While LLMs have shown remarkable capabilities in various educational applications (Becerra et al., 2024), their potential for addressing class imbalance in educational data mining remains largely unexplored. This gap is particularly notable in the context of MOOC forum analysis, where maintaining the authentic characteristics of student communications while generating synthetic data is crucial. The existing literature has approached the urgent post identification problem through various methods, from traditional machine learning approaches (Almatrafi et al., 2018) to sophisticated deep learning models (El-Rashidy et al., 2023). However, these approaches typically struggle with the inherent class imbalance in the data, leading to suboptimal performance in minority classes – often the most critical urgent posts. While data augmentation techniques have been explored in other domains, their application to MOOC forum data presents unique challenges due to the complex nature of educational discourse and the need to preserve semantic coherence. Building upon these challenges and opportunities, we explore three interconnected research questions:

- *RQ 1: How effective are LLM-generated synthetic posts in addressing class imbalance in MOOC forum urgency prediction?*
- *RQ 2: What is the impact of balanced training data on different machine learning models' performance across various urgency levels?*
- *RQ 3: To what extent does synthetic data augmentation affect model generalization and stability across different urgency categories?*
-

This work advances the field in several significant ways. We introduce the pipeline for using LLMs to generate balanced educational forum data while preserving semantic coherence and urgency characteristics,

provide empirical evidence of synthetic data augmentation's effectiveness in improving urgency prediction across multiple machine learning models, and offer comprehensive insights into the performance improvements and stability across urgency levels. These contributions not only advance our theoretical understanding of data augmentation in educational contexts but also provide practical solutions for improving MOOC forum management and learner support.

Related work

The challenge of urgent post detection and data imbalance in MOOCs

MOOC discussion forums, vital interaction hubs often analyzed for various predictive insights such as problem-solving depth (Li et al., 2023), frequently suffer information overload (Guo et al., 2019). Identifying posts requiring urgent instructor intervention is crucial (Almatrafi et al., 2018), with ongoing efforts focused on improving accuracy using methods like contextual features (El-Rashidy et al., 2024) or BERT-based semantic refinement (Zhang et al., 2024; El-Rashidy et al., 2023), as well as enhancing generalizability across different MOOCs (Švábenský et al., 2023). However, these efforts are persistently challenged by severe class imbalance. This imbalance, a known issue in educational data mining (Alrajhi et al., 2024), stems from the rarity of critical posts. Traditional sampling techniques often prove ineffective for complex educational discourse, failing to capture necessary linguistic and semantic patterns (Alrajhi et al., 2024), consequently hindering urgency detection accuracy (Sultani & Daneshpour, 2024). This necessitates novel approaches to address the data scarcity for critical, high-urgency posts.

LLM-based synthetic data generation

Large Language Models (LLMs) offer new possibilities for addressing educational data challenges (Becerra et al., 2024), including their potential for synthetic data generation to combat class imbalance, as pioneered in education by Li and Botelho (2024). Unlike generic data augmentation, generating high-fidelity educational data requires careful handling of contextual complexity and authenticity. While LLMs show promise in analyzing educational interactions (e.g., cognitive presence, Hu et al., 2024), creating realistic synthetic forum posts demands preserving not only linguistic style but also pedagogical context and relevance (Zhang et al., 2024). Maintaining authenticity in AI-generated educational content is paramount (Zobel & Meinel, 2024). Therefore, our work leverages LLMs specifically to generate a balanced dataset for urgency prediction, focusing on these quality requirements to overcome the limitations imposed by data imbalance.

Methodology

Our study utilizes the Stanford MOOCPosts dataset, a comprehensive collection of MOOC forum interactions with expert-annotated urgency levels ranging from 1 (not urgent) to 7 (highly urgent). The dataset exhibits significant class imbalance, with urgency level 1 containing 1,275 posts while level 7 has only 2 posts, presenting a clear challenge for traditional machine learning approaches.

Data generation

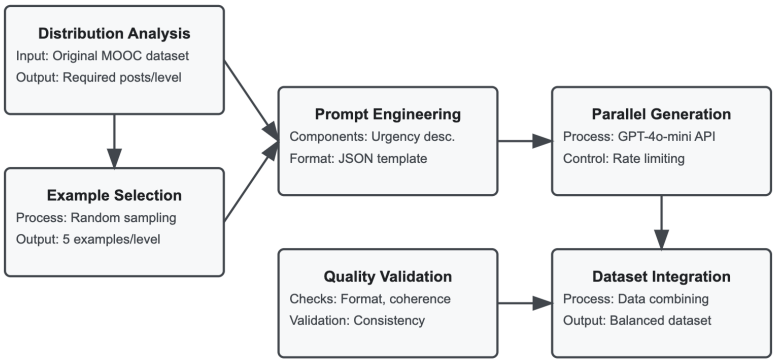
To address the class imbalance challenge, we developed a systematic data augmentation workflow (see in Figure 1). The workflow consists of six key stages, beginning with a distribution analysis of the original dataset to determine the required number of synthetic posts for each urgency level. The system then carefully selects representative examples and constructs prompts that capture the essential characteristics of each urgency level. These prompts guide the LLM in generating contextually appropriate synthetic posts through a parallel processing architecture.

Our generation process employs the GPT-4o-mini model, chosen for its balance of capability, API access, and cost-efficiency. Robust quality control mechanisms were integrated into the generation process. Automated validation after each API call ensured the generated output adhered to the required JSON format, contained all necessary fields, met a minimum post length requirement, and reported an urgency level within the valid 1-7 range. Consistency and semantic coherence were primarily ensured through careful prompt engineering: the prompts provided the LLM (GPT-4o-mini) with explicit urgency level definitions, multiple representative examples from the original dataset for the target level, and detailed instructions to guide the generation of contextually appropriate posts. This careful attention to quality ensures that synthetic posts maintain the authentic characteristics of MOOC forum interactions while addressing the distributional imbalance. The workflow successfully generated and validated synthetic posts to achieve a balanced dataset of exactly 1,275 posts per

urgency level. This required varying numbers of synthetic posts for each level: 55 for level 2, 941 for level 3, 902 for level 4, 1,025 for level 5, 1,227 for level 6, and 1,273 for level 7.

Figure 1

LLM-Based Data Generation Workflow



This workflow outlines the key stages for creating a balanced dataset: analyzing distribution, selecting examples, prompt engineering, LLM generation (GPT-4o-mini), quality validation, and data integration.

Model training and evaluation

Our evaluation workflow implements a comprehensive pipeline that combines sophisticated text preprocessing with multiple machine learning approaches. The preprocessing stage utilizes TF-IDF vectorization with a maximum of 1,000 features and English stop word removal, ensuring efficient yet comprehensive text representation. To assess augmentation impact across diverse algorithms, we selected five models: Decision Tree (baseline), common ensembles (Random Forest, XGBoost, Gradient Boosting), and SVR (effective in high dimensions). These models were each configured with carefully tuned hyperparameters for optimal performance. For evaluation, the balanced dataset, combining original posts and all LLM-generated synthetic posts, was created upfront. Subsequently, a standard 5-fold cross-validation strategy was applied to the training portion of this combined dataset using cross-validation to tune and assess model robustness. At the same time, final performance was measured on a held-out test set also drawn from the balanced data. This process utilized multiple scoring metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2. Our workflow calculates both validation and test scores, providing a comprehensive view of model generalization. Performance analysis is conducted at both the aggregate and per-urgency-level granularity, enabling a detailed understanding of how data augmentation affects prediction accuracy across different urgency categories. This methodological approach ensures reproducibility and scalability while providing valuable insights into the effectiveness of LLM-based data augmentation in educational contexts.

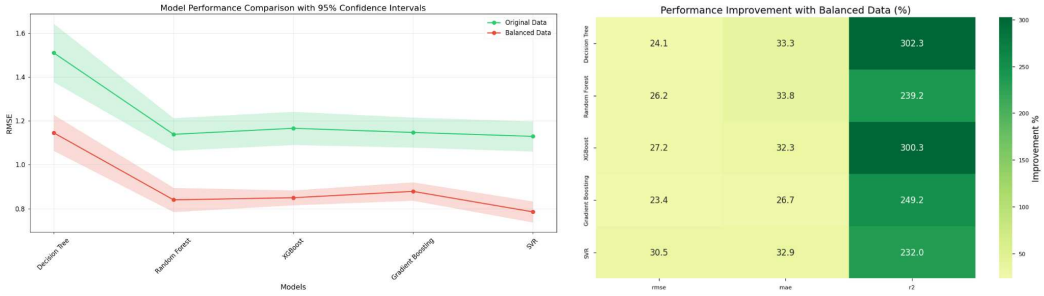
Results and analysis

Our experimental results demonstrate substantial improvements in model performance through LLM-based data augmentation. The comprehensive evaluation reveals benefits in both prediction accuracy and model stability across different urgency levels.

The performance comparison with confidence intervals demonstrates significant improvements in model predictions after applying our data augmentation approach (see Figure 2). Particularly noteworthy is the Support Vector Regression (SVR) model's enhanced performance, achieving a reduction in validation RMSE from 1.129 to 0.785. The narrowed confidence intervals across all models indicate increased prediction stability, suggesting that the synthetic data effectively captures the essential characteristics of MOOC forum posts.

Figure 2

Evaluation Comparing Model Performance Before and After Augmentation



Left: Validation RMSE (lower is better) with 95% CIs for models on original (green) vs. balanced (red) data. Right: Percentage improvement (%) in key metrics (RMSE, MAE, R^2) when using balanced data.

The visualization of performance improvements reveals substantial gains across multiple evaluation metrics (see Figure 2). The most dramatic improvements appear in R^2 scores, with Decision Trees showing over 400% improvement and ensemble methods achieving more than 200% enhancement. This significant improvement in R^2 values, coupled with substantial reductions in both RMSE and MAE, indicates that our balanced dataset enables more accurate predictions across the full spectrum of urgency levels.

The analysis of high-urgency predictions reveals particularly promising results for critical cases that demand immediate instructor attention (see Table 1). The dramatic reduction in RMSE for urgency levels 5 and 6, coupled with newly enabled predictions for level 7, demonstrates our approach's effectiveness in addressing the crucial challenge of identifying highly urgent posts. The balanced dataset notably improved the models' ability to identify these rare but critical cases, with RMSE improvements of up to 80% for level 6 posts.

Table 1

Model Performance on High-Urgency Posts

Urgency Level	Original RMSE	Balanced RMSE	Improvement
Level 5	2.193	0.946	56.9%
Level 6	3.537	0.708	80.0%
Level 7	N/A*	0.283	N/A

* Based on N=2 original samples of Level 7, interpretation requires caution.

Error analysis reveals remaining challenges primarily in posts containing multiple urgency aspects or subtle contextual indicators. These findings suggest that while our approach successfully addresses the fundamental challenge of class imbalance, future work might explore more sophisticated methods for handling complex, multi-faceted forum posts.

Discussion

Our study advances LLM-based data augmentation in education by extending beyond Li and Botelho's (2024) binary classification of at-risk students to a seven-level urgency detection pipeline. This approach enhances LLM-generated synthetic data capabilities, achieving an 80% RMSE reduction in high-urgency predictions (level 6), and effectively identifying posts requiring immediate intervention (El-Rashidy et al., 2024). Unlike traditional augmentation methods (Alrajhi et al., 2024), our method preserves semantic coherence while addressing class imbalance in educational contexts.

Several limitations warrant careful consideration. While our synthetic data generation process effectively balances the dataset, the reliance on GPT-4o-mini may introduce model-specific biases. The current approach focuses solely on textual content without incorporating temporal patterns or user interaction histories that might provide additional context for urgency assessment. Also, while performing data augmentation upfront facilitated consistent model comparison, this procedural choice warrants consideration regarding potential data leakage prior to cross-validation. Moreover, while our evaluation metrics show significant statistical improvements, the practical impact on instructor response times and student outcomes requires further investigation to fully validate the approach's educational value.

Future research could explore more sophisticated approaches to synthetic data generation, potentially incorporating multi-modal features and temporal dynamics to better capture the complexity of MOOC interactions. Additionally, expanding this approach to cross-platform and cross-domain applications could establish its broader applicability in diverse educational contexts, potentially leading to more robust and generalizable urgency detection systems.

References

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015, February 9). YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. National Science Foundation. http://ilpubs.stanford.edu:8090/1120/

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education*, 118, 1–9. https://doi.org/10.1016/j.compedu.2017.11.002

Becerra, Á., Mohseni, Z., Sanz, J., & Cobos, R. (2024). A generative AI-based personalized guidance tool for enhancing the feedback to MOOC learners. In *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–8). IEEE. https://doi.org/10.1109/EDUCON60312.2024.10578809

El-Rashidy, M. A., Khodeir, N. A., Farouk, A., Aslan, H. K., & El-Fishawy, N. A. (2024). Attention-based contextual local and global features for urgent posts classification in MOOCs discussion forums. *Ain Shams Engineering Journal*, 15(4), 102605. https://doi.org/10.1016/j.asej.2023.102605

El-Rashidy, M. A., Farouk, A., El-Fishawy, N. A., Aslan, H. K., & Khodeir, N. A. (2023). New weighted BERT features and multi-CNN models to enhance the performance of MOOC posts classification. *Neural Computing and Applications, 35(*24), 18019–18033. https://doi.org/10.1007/s00521-023-08673-z

Guo, S. X., Sun, X., Wang, S. X., Gao, Y., & Feng, J. (2019). Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums. IEEE Access, 7, 120522–120532. https://doi.org/10.1109/ACCESS.2019.2929211

Hu, Y., Giacaman, N., & Donald, C. (2024). Enhancing Trust in Generative AI: Investigating Explainability of LLMs to Analyse Confusion in MOOC Discussions.

Li, H., & Botelho, A. F. (2024). Fine-tuning large language models for data augmentation to detect at-risk students in online learning communities. In 17th *International Conference on Computer-Supported Collaborat*ive Learning (CSCL) (pp. 441–442). ISLS. https://doi.org/10.22318/cscl2024.208036

Li, J., Li, L., Zhu, Z., & Shadiev, R. (2023). Research on the predictive model based on the depth of problem-solving discussion in MOOC forum. Education and Information Technologies, 28(10), 13053-13076.

Stanford University. (2014). The Stanford MOOCPosts Data Set. Stanford.edu. https://datastage.stanford.edu/StanfordMoocPosts/#procedures

Sultani, M., & Daneshpour, N. (2024). Extracting urgent questions from MOOC discussions: A BERT-based multi-output classification approach. Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-024-09090-7

Valdemar Švábenský, Ryan Baker, Andrés Zambrano, Yishan Zou, & Stefan Slater. (2023). Towards Generalizable Detection of Urgency of Discussion Forum Posts. Proceedings of the 16th International Conference on Educational Data Mining, 302--309. https://doi.org/10.5281/zenodo.8115790

Zhang, Z., Zhu, X., He, Q., & Zhang, L. (2024). BERT-based global semantic refinement and local semantic extraction for distinguishing urgent posts in MOOC forums. IEEE Access, 12, 116250–116258. https://doi.org/10.1109/ACCESS.2024.3426976

Zobel, T., & Meinel, C. (2024). Comparing AI in online learning: The transition and trade-offs between intent-based learning assistants and LLM-chatbots in MOOCs. In *2024 IEEE Digital Education and MOOCs Conference (DEMOcon)* (pp. 1–6). IEEE. https://doi.org/10.1109/DEMOcon63027.2024.10748211