International Society of the Learning Sciences

Fine-Tuning Large Language Models for Data Augmentation to Detect At-Risk Students in Online Learning Communities

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

Abstract: We introduce a working approach that combines the method of fine-tuning large language models (LLMs) to create augmented data for the regression predictive models aimed at detecting at-risk students in online learning communities. This approach has the potential to leverage scarce data to improve urgency detection, and it can also present the role of artificial intelligence in enhancing the resilience of educational communities and ensuring timely interventions within online learning settings.

Introduction

Online learning communities, epitomized by Massive Open Online Courses (MOOCs), have become pivotal in fostering accessible education. Yet, identifying and supporting at-risk students in these environments remains a complex challenge (Romero et al., 2010; Marbouti et al., 2016). Prior research has applied machine learning to predict student performance and dropout in MOOCs using features like forum posts, assignment submissions, and quiz scores (Kloft et al., 2014). The use of text as a data source for urgency detection in educational settings adds to this body of work and is related to sentiment analysis (Nasukawa & Yi, 2003).

Large Language Models (LLMs), such as GPT-3, have been extensively utilized for natural language processing tasks (Brown et al., 2020). However, their application in educational settings, particularly for the detection of at-risk students, is less explored. The use of LLMs for data augmentation, as proposed in this study, intersects with ongoing research into overcoming data scarcity in machine learning (Shorten et al., 2019). XGBoost, an advanced gradient-boosting machine learning algorithm, has been successfully applied in various domains, including educational data mining (Chen et al., 2013). In the context of resilience and responsiveness of educational communities, AI has played a role in adaptive systems that personalize the educational experience based on student needs (Pechenizkiy et al., 2009).

Building upon these foundations, our work extends the field by employing fine-tuned LLMs for data augmentation to enhance the predictive modeling of at-risk students. This innovative approach prompts the following research questions: *RQ1***:** How can fine-tuned LLMs be utilized to generate synthetic data that improves the accuracy of urgency detection in MOOCs? *RQ2***:** To what extent does training the XGBoost model with the augmented dataset generated by fine-tuned LLMs enhance the model's performance in identifying at-risk students compared to training with the original dataset?

Methodology and findings

The cornerstone of our methodology is the Stanford MOOCPosts dataset (Stanford University, 2014), containing 29,604 anonymized posts from public online courses, with a subset of 3,505 entries being used for this study. These entries have been manually scored for urgency on a scale from 1 (least urgent) to 7 (most urgent) to reflect the need for immediate attention from the educational institution. We developed a model using the XGBoost regression algorithm and applied the following: *Preprocessing:* Utilize standardized NLP procedures to normalize, tokenize, remove stop words, and perform stemming on text data to ensure uniformity and reduce noise. *Feature Extraction:* TF-IDF Vectorization was applied to convert the processed text into a feature matrix, which captures the importance of terms within the documents. *Model Training:* The XGBoost model was trained on the original dataset, with data split into training (70%), validation (10%), and testing (20%) sets.

For fine-tuning, we used GPT-3.5turbo and considered the open-source LLaMA2 model as an alternative due to its cost-effectiveness for future applications. The fine-tuning process involves adjusting the models based on our urgency-scored dataset, using prompt engineering to guide the LLMs. We transform the initial data into a format suitable for LLM processing, as exemplified by the script provided, which restructures data into a JSON-lines (jsonl) format that defines the role and content for system-user-assistant interactions. This clarifies the task and expected output for the model, allowing it to generate posts that mimic the style and urgency of the students' original content. After the initial training, additional synthetic data is created using the fine-tuned LLM to simulate the distribution of urgency in student posts, doubling the size of data. The models' performance was evaluated using the test set (20%) from the original dataset to establish a baseline.

The analysis of our model's performance, as detailed in Table 1, indicates a modest enhancement in predictive accuracy following the introduction of augmented data generated through fine-tuned large language models. While the decrease in Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) for the augmented dataset model signals a positive trend, the improvements are small.

This incremental progress underscores the nuanced nature of urgency detection within educational texts and suggests that further refinement of the data generation process and predictive modeling techniques is necessary. The modest gains prompt more profound consideration of the complexities in interpreting student discourse and point to additional layers of sophistication in our Fine-tuned LLM.

Table 1

Model Performance Before and After Using Fine-Tuned LLM Generation for Augmented Data

	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error
Original Dataset	1.370	1.170	0.911
Augmented Dataset	1.299	1.140	0.877

This study introduces an innovative AI-based approach for identifying at-risk students in MOOCs by fine-tuning LLMs to generate synthetic data for enhancing predictive models. The preliminary XGBoost model, trained on a dataset with human-annotated urgency scores, provides insights for more sophisticated AI applications in online learning. While initial results are promising, as evidenced by modest improvements in MSE, RMSE, and MAE (Table 1), further refinement is needed. The incremental gains highlight the complexities of detecting urgency in educational discourse (Nasukawa & Yi, 2003) and align with research on overcoming data scarcity in machine learning (Shorten et al., 2019).

This research underscores the critical role of data quality and volume in training effective AI systems for educational applications. It also provides insights into the scalability of AI solutions, demonstrating how open-source models can be leveraged to achieve outcomes comparable to their resource-intensive counterparts. Our findings contribute to the ongoing discourse on the practical challenges and ethical implications of implementing AI in education, including cost, data privacy, and the use of AI-generated content. Should AI-augmented data prove effective, it would underscore the viability of LLMs in educational technology.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. Computational and Mathematical Methods in Medicine, 2013, e873595. https://doi.org/10.1155/2013/873595

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In Proc. of the EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs (pp. 60-65).

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. P. C. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, 103, 1-15.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing Favorability Using Natural Language Processing. Proceedings of the International Conference on Knowledge Capture - K-CAP '03. https://doi.org/10.1145/945645.945658

Pechenizkiy, M., Trcka, N., Vasilyeva, E., & van der Aalst, W. (2009). Process mining online assessment data. In Proc. of the 2nd Int'l Conference on EDM (pp. 279-288).

Romero, C., Ventura, S., & Garcia, E. (2010). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2019). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Molecular pharmaceutics, 16(7), 2776-2790.

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

Acknowledgments

We would like to thank Dr. Jinnie Shin from the College of Education at the UF, NSF (e.g., 2331379, 1903304, 1822830, 1724889), as well as IES (R305B230007), and MathNet.