Integrating Optuna And Synthetic Data Generation For Optimized Medical Transcript Classification Using BioBERT

Authors : Sachi Nandan Mohanty, Shreya Sinha, Sweeti Sah, Shweta Sharma

Abstract : The advancement of natural language processing has majorly influenced the field of medical transcript classification, providing a robust framework for enhancing the accuracy of clinical data processing. It has enormous potential to transform healthcare and improve people's livelihoods. This research focuses on improving the accuracy of medical transcript categorization using Bidirectional Encoder Representations from Transformers (BERT) and its specialized variants, including BioBERT, ClinicalBERT, SciBERT, and BlueBERT. The experimental work employs Optuna, an optimization framework, for hyperparameter tuning to identify the most effective variant, concluding that BioBERT yields the best performance. Furthermore, various optimizers, including Adam, RMSprop, and Layerwise adaptive large batch optimization (LAMB), were evaluated alongside BERT's default AdamW optimizer. The findings show that the LAMB optimizer achieves equally good performance as AdamW. Synthetic data generation techniques from Gretel were utilized to augment the dataset, expanding the original dataset from 5,000 to 10,000 rows. Subsequent evaluations demonstrated that the model maintained its performance with synthetic data, with the LAMB optimizer showing marginally better results. The enhanced dataset and optimized model configurations improved classification accuracy, showcasing the efficacy of the BioBERT variant and the LAMB optimizer. It resulted in an accuracy of up to 98.2% and 90.8% for the original and combined datasets, respectively.

Keywords : BioBERT, clinical data, healthcare AI, transformer models
Conference Title : ICCAIP 2025 : International Conference on Computer Analysis of Images and Patterns
Conference Location : Paris, France
Conference Dates : May 08-09, 2025