Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

University of Westminster, Coat of Arms

SyntheText

Towards Universal NLP: Explainable AI-Guided Noise Injection for Language Agnostic Automated Text Data Augmentation

A dissertation by

Mr. Wijesundera Arachchige Sarith Manthusa Wijesundera

w1912785 | 20210010

Supervised by

Mr. Suresh Peiris

April 2025

Submitted in partial fulfilment of the requirements for the

BEng (Hons) Software Engineering degree at the University of Westminster.

ABSTRACT

Data augmentation is a powerful strategy to address data scarcity. It improves the generalisation and robustness of machine learning models which mitigates model overfitting issues. Data augmentation is in its early stages in natural language processing due to the difficulty in textual data transformation. Therefore, selecting the optimal augmentation techniques has become a challenging task while preserving keywords of texts. In the context of multiple languages, challenges arise because an augmentation technique that performs well in one language may not perform the same in another language.

The author proposes a novel approach to automate the pipeline of textual data augmentation by enabling cross-lingual capabilities through the use of language models and noising-based augmentation strategies. By framing the problem as a hyperparameter optimisation task, the author defines the augmentation search space with language agnostic noise injection techniques and a restructured augmentation policy. To maintain keywords of texts during the augmentation, XAI techniques are leveraged to compute the word contribution scores on the model prediction.

Experiments were conducted for sentiment analysis in three languages: English, Sinhala, and Korean, within an extremely low-resource setting, utilising only 80 training samples and 60 validation samples. XLM-R-base was used as the backbone classifier model. For English, the accuracy improvement after applying the proposed augmentation strategies increased by 16%-18%, while for Sinhala, the improvement was 9%-10% and for Korean, it ranged from 21%-23%. These results highlight the potential of leveraging noising-based techniques and XAI to perform language agnostic automated data augmentation in low-resource contexts.

Keywords: Automated Text Data Augmentation, Low-resource Languages, Cross-lingual Transfer Learning, Hyperparameter Optimisation, Explainable AI

Subject Descriptors:

- Computing methodologies → Artificial intelligence → Natural language processing → Natural language generation
- Computing methodologies → Machine learning → Machine learning algorithms → Regularisation