ABSTRACT

THE EFFECT OF DATA AUGMENTATION ON THE PERFORMANCE OF PRE-TRAINED TRANSFORMER MODELS IN COVID-19 NEWS TITLE CLASSIFICATION

By

YULINA PUTRI

The volume of digital information, especially online news during the COVID-19 pandemic has surged, making it difficult for readers to obtain relevant information. Automatic classification based on pre-trained transformer models is an approach that can be used to filter information efficiently. However, the performance of classification models is greatly affected by the quality and distribution of training data. The imbalance of class distribution in the data can degrade the model's prediction performance, especially for minority classes. Therefore, this study aims to analyze the effect of data augmentation on improving the performance of classification models on imbalanced COVID-19 news title data. Two augmentation methods are used, namely synonym replacement (SR) and a combination of synonym replacement with back translation (SR + BT). The evaluation was carried out using several pre-trained transformer models, including ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, and RoBERTa. The results show that the augmentation successfully improved the average accuracy and recall up to 1.45%, precision and F1-score up to 1.94%, and ROC-AUC up to 5.58%. In the minority class, augmentation increased recall by 14.67% and F1-score by 5.37%. This finding shows that data augmentation is effective in overcoming class imbalance and improving the model's ability to identify minority classes that were previously difficult to detect.

Keywords: Classification, Transformer, Data Augmentation, Online News, Synonym Replacement, Back Translation.

ABSTRAK

PENGARUH AUGMENTASI DATA TERHADAP KINERJA *PRE-TRAINED TRANSFORMER MODELS* DALAM KLASIFIKASI JUDUL BERITA COVID-19

Oleh

YULINA PUTRI

Volume informasi digital, khususnya berita *online* selama masa pandemi COVID-19 mengalami lonjakan yang menyebabkan pembaca kesulitan memperoleh informasi yang relevan. Klasifikasi otomatis berbasis *pre-trained transformer models* merupakan pendekatan yang dapat digunakan untuk menyaring informasi secara efisien. Namun, kinerja model klasifikasi sangat dipengaruhi oleh kualitas dan distribusi data latih. Ketidakseimbangan distribusi kelas dalam data dapat menurunkan kinerja prediksi model, terutama terhadap kelas minoritas. Oleh karena itu, penelitian ini bertujuan menganalisis pengaruh augmentasi data terhadap peningkatan kinerja model klasifikasi pada data judul berita COVID-19 yang tidak seimbang. Dua metode augmentasi yang digunakan yaitu *synonym replacement* (SR) dan kombinasi *synonym replacement* dengan *back translation* (SR + BT). Evaluasi dilakukan menggunakan beberapa model *pre-trained transformer*, diantaranya ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa. Hasil menunjukkan bahwa augmentasi berhasil meningkatkan rata-rata akurasi dan *recall* hingga 1,45%, presisi dan *F1-score* hingga 1,94%, dan ROC-AUC hingga 5,58%. Pada kelas minoritas, augmentasi meningkatkan *recall* hingga 14,67% dan *F1-score* hingga 5,37%. Temuan ini menunjukkan bahwa augmentasi data efektif dalam mengatasi ketidakseimbangan kelas dan meningkatkan kemampuan model dalam mengidentifikasi kelas minoritas yang sebelumnya sulit terdeteksi.

Kata-kata kunci: Klasifikasi, *Transformer*, Augmentasi Data, Berita *Online*, *Synonym Replacement*, *Back Translation*.