INTISARI

OPTIMASI INDOBERT UNTUK ANALISIS SENTIMEN PADA DATA TIDAK SEIMBANG DENGAN AUGMENTASI DATA : STUDI KASUS KEBIJAKAN TAPERA DAN DATASET BENCHMARK

Oleh

Dina Febriana

24/546904/PPA/06861

Ketidakseimbangan data dapat menyebabkan model IndoBERT cenderung mempelajari kelas mayoritas, sehingga nilai presisi dan *recall* pada kelas minoritas menjadi lebih rendah meskipun akurasi keseluruhan relatif tinggi. Oleh karena itu, diperlukan penerapan metode augmentasi data untuk meningkatkan keragaman dan representasi data pada kelas minoritas. Penelitian ini menerapkan beberapa metode augmentasi data, yaitu *Synonym Replacement*, *Back-Translation*, *Masked Language Modeling* (MLM), serta kombinasi MLM dengan *Cosine Similarity*. Seluruh metode augmentasi, kecuali *Back-Translation* diterapkan dengan rasio penggantian kata sebesar 15%, 25%, dan 35%.

Model dikonfigurasikan dengan kombinasi hyperparameter terbaik yang diperoleh adalah jumlah *epoch* sebesar 20, *batch size* sebesar 8, *learning rate* sebesar 0.00002, *warmup steps* sebanyak 400, *weight decay* sebesar 0.01, *early stopping* sebesar 5, *dropout rate* sebesar 0.25 serta fungsi loss menggunakan *CrossEntropyLoss*.

Hasil penelitian menunjukkan bahwa pada model IndoBERT, dataset yang digunakan cukup representatif untuk proses *fine-tuning* meskipun memiliki ketidakseimbangan data. Hal ini ditunjukkan oleh performa model yang lebih optimal ketika dilatih menggunakan data asli tanpa augmentasi dibandingkan dengan data hasil augmentasi. Selain itu, pada model IndoBERTweet performa terbaik pada dataset kebijakan tapera diperoleh melalui augmentasi MLM dengan rasio 25%, sedangkan pada dataset benchmark performa optimal dicapai melalui augmentasi *synonym replacement* dengan rasio 25%.

Kata Kunci: IndoBERT, IndoBERTweer, Augmentasi Data, MLM.

ABSTRACT

OPTIMIZING INDOBERT FOR SENTIMENT ANALYSIS ON IMBALANCED DATA USING DATA AUGMENTATION : A CASE STUDY ON THE TAPERA POLICY AND BENCHMARK DATASETS

By

Dina Febriana

24/546904/PPA/06861

Data imbalance can cause the IndoBERT model to be biased toward the majority class, resulting in lower precision and recall values for minority classes, even though the overall accuracy remains relatively high. Therefore, data augmentation methods are required to enhance the diversity and representation of minority class data. This study applies several data augmentation techniques, namely Synonym Replacement, Back-Translation, Masked Language Modeling (MLM), and a combination of MLM with Cosine Similarity. All augmentation methods, except Back-Translation, are applied with word replacement ratios of 15%, 25%, and 35%.

The model is configured using the best performing hyperparameter combination, consisting of 20 epochs, a batch size of 8, a learning rate of 0.00002, 400 warmup steps, a weight decay of 0.01, early stopping of 5, a dropout rate of 0.25, and CrossEntropyLoss as the loss function.

The experimental results indicate that, for the IndoBERT model, the datasets used are sufficiently representative for the fine-tuning process despite data imbalance. This is demonstrated by the superior performance achieved when the modal is trained using the original dataset without augmentation compared to augmented datasets. Furthermore, for the IndoBERTweet model, the best performance on the tapera policy dataset is obtained using MLM based augmentation with a 25% rasio, while on the benchmark dataset, optimal performance is achieved using synonym replacement with a 25% ratio.

Keywords: IndoBERT, IndoBERTweet, Data Augmentation, MLM.