ABSTRACT

GENERATED TEXT IN RESTRUCTURING AN IMBALANCED DATASET TITLE COVID-19 NEWS USING DEEP LEARNING MODEL

By

FEBI SITI SUTRIA NINGSIH

Imbalanced data is a problem that often occurs in classifying. This has the potential to cause bias against the majority class. Oversampling or text generation is one of the methods used to overcome imbalanced data. This study proposes the creation of a synonym-based text to restructure the imbalanced covid-19 online news dataset. Deep learning models (CNN and LSTM) use the word embedding feature and validated using the k-fold cross validation method for imbalanced and balanced data. The results showed that using both imbalanced and balanced data, CNN achieved significantly higher performance by 7% - 9% in accuracy, precision, recall, and f1-score than LSTM. Although the text generation method is proven to be able to overcome the problem of data imbalance, the text generation method has a drawback, namely that there is often a loss of semantic meaning.

Keywords: Text Generation, Imbalanced Dataset, Deep Learning, k-Fold Cross Validation.

ABSTRAK

GENERATED TEXT DALAM RESTRUKTURISASI KETIDAKSEIMBANGAN DATASET JUDUL BERITA *COVID-19* DENGAN MENGGUNAKAN MODEL *DEEP LEARNING*

Oleh

FEBI SITI SUTRIA NINGSIH

Imbalanced data merupakan suatu permasalahan yang sering terjadi dalam melakukan klasifikasi. Hal ini berpotensi menyebabkan bias terhadap kelas mayoritas. *Oversampling* atau *text generation* adalah salah satu metode yang digunakan untuk mengatasi *imbalanced* data. Penelitian ini mengusulkan pembuatan teks berbasis sinonim untuk merestrukturisasi dataset berita *online covid-19* yang tidak seimbang. Model *deep learning* (CNN dan LSTM) yang menggunakan fitur *word embedding* dan divalidasi menggunakan metode *k-fold cross validation* untuk *imbalanced* dan *balanced* data. Hasil penelitian menunjukkan bahwa baik dengan menggunakan data *imbalanced* dan *balanced* data, CNN mencapai kinerja yang secara signifikan lebih tinggi sebesar 7% - 9% dalam akurasi, presisi, *recall*, dan *f1-score* daripada LSTM. Meskipun metode *text generation* terbukti dapat mengatasi masalah ketidakseimbangan data, tetapi metode *text generation* memiliki kekurangan yaitu sering terjadi kehilangan makna semantik.

Kata kunci: *Text Generation*, Ketidakseimbangan Data, *Deep Learning*, *k-Fold Cross Validation.*