INTISARI

KOMBINASI DATA AUGMENTASI DAN SKEMA TERM WEIGHTING UNTUK ANALISIS SENTIMEN

Oleh:

Very Dwi Setiawan
21/484935/PPA/01699

Analisis sentimen adalah teknik yang digunakan untuk mengenali dan memahami sentimen atau opini yang terkandung dalam sebuah teks. Salah satu masalah yang dihadapi dalam analisis sentimen adalah *data imbalance*, di mana jumlah sampel positif, negatif, dan netral tidak seimbang. *Data imbalance* dapat mengurangi kinerja model klasifikasi sentimen. Untuk mengatasi masalah ini, penelitian ini mengusulkan metode untuk mengatasi *data imbalance* menggunakan teknik *easy data augmentation* (EDA). Kemudian teknik EDA dikombinasikan dengan skema *term weighting* TF-IDF, TF-RF, TF-IDF-ICF, dan TF-IDF-ICSDF. Metode yang diusulkan diterapkan pada *dataset* IndoNLU, IGSA, Covid-19, dan PPKM. Di mana masing-masing *dataset* menunjukkan rasio *data imbalance* yang berbeda. Tujuan yang ingin dicapai dalam penelitian ini adalah meningkatkan akurasi model dengan kombinasi teknik *data augmentation* dan skema *term weighting.* Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi *data augmentation* dan skema *term weighting* memberikan peningkatan kinerja yang baik untuk SVM maupun *Random Forest*. Kombinasi EDA *Random Swap* dan TF-IDF-ICF menghasilkan nilai akurasi tertinggi pada pemodelan SVM, yaitu 91,23%, dengan peningkatan sebesar 5,24%. Sedangkan pada pemodelan *Random Forest*, akurasi tertinggi ditunjukkan pada kombinasi EDA *Random Swap* dan TF-IDF-ICSDF, yaitu 93,92%, dengan peningkatan sebesar 11,57%.

Kata Kunci: Analisis Sentimen, EDA, Skema *Term Weighting, Random Forest,* SVM

ABSTRACT

COMBINATION OF DATA AUGMENTATION AND TERM WEIGHTING SCHEME FOR SENTIMENT ANALYSIS

Oleh:

Very Dwi Setiawan
21/484935/PPA/01699

Sentiment analysis is a technique used to recognize and understand the sentiments or opinions contained in a text. One of the problems encountered in sentiment analysis is data imbalance, in which the number of positive, negative and neutral samples is unequal. Data imbalance can reduce the performance of sentiment classification models. To overcome this problem, this study proposes a method to overcome data imbalance using easy data augmentation (EDA) techniques. Then the EDA technique is combined with the TF-IDF, TF-RF, TF-IDF-ICF, and TF-IDF-ICSDF term weighting schemes. The proposed method the researchers apply to the IndoNLU, IGSA, Covid-19, and PPKM *dataset*. Where each *dataset* shows a different data imbalance ratio. The goal the researchers achieve in this study is to increase the model's accuracy with a combination of data augmentation techniques and term weighting schemes. Overall, the results of this study indicate that the combination of augmentation data and term weighting schemes provides a good performance increase for both SVM and Random Forest. The combination of EDA Random Swap and TF-IDF-ICF produced the highest accuracy value in SVM modeling, namely 91.23%, with an increase of 5.24%. In the Random Forest modeling, the highest accuracy the results shown in the combination of EDA Random Swap and TF-IDF-ICSDF, which is 93.92%, with an increase of 11.57%.

Keywords: Sentiment analysis, EDA, Term Weighting Scheme, Random Forest, SVM