INTISARI

Imbalanced dataset merupakan permasalahan yang sering ditemukan dalam proses penelitian tentang klasifikasi. Data dalam kondisi imbalance mempengaruhi tingkat keakuratan prediksi model seperti yang terjadi pada klasifikasi komentar program Kampus Merdeka yang peneliti lakukan. Penelitian ini terfokus pada penanganan imbalanced dataset untuk meningkatkan kinerja klasifikasi komentar yang berasal dari aplikasi Twitter. Data komentar diklasifikasikan ke dalam empat kelas yaitu informasi, opini, pertanyaan, out of topic. Metode yang digunakan untuk balancing dataset adalah Near Miss, SMOTE, ADASYN, dan Random Combination Sampling. Evaluasi performa dilakukan menggunakan algoritma Support Vector Machine (SVM) dengan perbandingan komposisi data training dan testing 70:30, 80:20, dan 90:10. Melalui pengujian yang dilakukan, hasil terbaik diperoleh pada komposisi 90:10. Hal ini dapat dipahami karena mesin cenderung belajar dan berlatih dengan data yang lebih banyak daripada komposisi lainnya. Hasil yang diperoleh pada nilai akurasi, F1-Score, dan kurva ROC-AUC menunjukkan hasil yang serupa. Hasil tertinggi diperoleh pada metode ADASYN dengan nilai F1-Score 0,9. Sedangkan hasil terendah diperoleh pada metode Near Miss dengan nilai F1-Score 0,68. Secara umum, dapat disimpulkan bahwa metode balancing dataset yang digunakan telah diimplementasikan sesuai prosedur dan menghasilkan peningkatan kinerja model yang cukup memuaskan.

Kata kunci: Imbalanced Dataset, Kampus Merdeka, Support Vector Machine, Twitter, Klasifikasi Teks

ABSTRACT

Imbalanced dataset is a problem that is often found in classification. An imbalanced condition affects the level of accuracy of model predictions as happened in the classification of the Kampus Merdeka program comments. This research focuses on handling the imbalanced dataset to improve the performance of the classification of comments from Twitter. The methods used are Near Miss, SMOTE, ADASYN, and Random Combination Sampling. Performance evaluation was carried out using the Support Vector Machine (SVM) algorithm with a composition of the training and testing data at 70:30, 80:20, and 90:10. Through the tests carried out, the best results were obtained with a composition of 90:10. This is understandable because machines tend to learn and train with more data. The results obtained on the accuracy value, F1-Score, and the ROC-AUC curve show similar results. The highest results were obtained in the ADASYN method with an F1-Score of 0.9. While the lowest results were obtained in the Near Miss method with an F1-Score value of 0.68. This can be concluded that the dataset balancing method used has been implemented according to procedures and resulted in a satisfactory increase in model performance.

Keywords: *Imbalanced Dataset, Kampus Merdeka, Support Vector Machine, Twitter, Text Classification*