ABSTRAK

Analisis sentimen merupakan salah satu cabang pengolahan bahasa alami yang bertujuan untuk mengidentifikasi polaritas emosi dalam teks. Bahasa Dayak Jalai, sebagai salah satu bahasa daerah di Indonesia, masih minim kajian dalam konteks analisis sentimen, termasuk oleh NusaX sebagai penyedia dataset multibahasa, sehingga penelitian ini bertujuan untuk mengembangkan model analisis sentimen berbasis Bahasa Dayak Jalai. Penelitian ini membahas analisis sentimen dalam Bahasa Dayak Jalai menggunakan data yang didapat dari NusaX. Dataset berjumlah 1000 data berbahasa Indonesia dengan label sentimen positif, negatif, dan netral yang kemudian ditransliterasi ke Bahasa Dayak Jalai. Setelah proses transliterasi, data melalui tahap *preprocessing*, dan pembobotan menggunakan *TF-IDF*. Analisis sentimen dilakukan menggunakan dua metode klasifikasi, yaitu Support Vector Machine (SVM) dan Random Forest (RF) dengan validasi silang K-Fold dengan nilai k = 3, 5, 7, dan 10. Pada metode SVM, digunakan kernel linear dan RBF, sedangkan pada Random Forest dilakukan *tuning* parameter meliputi *n_estimators*, *max_depth*, dan *min_samples_split*.

Hasil penelitian didapatkan bahwa metode SVM dengan kernel RBF memberikan performa terbaik pada validasi K-Fold 10 dengan parameter *C* = 7 dan *gamma* = 0.001, menghasilkan akurasi 81.30 %, presisi 82.21 %, *recall* 81.30 %, dan *F1-score* 81.27 %. Sementara itu, metode Random Forest memberikan performa terbaik pada validasi K-Fold 7 dengan parameter *n_estimators* = 100, *max_depth* = 40, dan *min_samples_split* = 2, menghasilkan akurasi 78.70 %, presisi 78.98 %, *recall* 78.70 %, dan *F1-score* 78.63 %.

Kata Kunci: Analisis Sentimen, nusaX, Bahasa Dayak Jalai, SVM, Random Forest, K-Fold.

ABSTRACT

Sentiment analysis is a branch of natural language processing aimed at identifying the emotional polarity in a text. Dayak Jalai, as one of Indonesia's regional languages, has not yet been studied in the context of sentiment analysis, including by NusaX as a provider of multilingual datasets. This research aims to develop a sentiment analysis model for the Dayak Jalai language. The study utilizes data obtained from NusaX, consisting of 1000 Indonesian-language entries labeled with positive, negative, and neutral sentiments, which were transliterated into the Dayak Jalai language. After transliteration, the data underwent preprocessing and weighting using TF-IDF. Sentiment analysis was conducted using two classification methods, namely Support Vector Machine (SVM) and Random Forest (RF), with K-Fold cross-validation (k = 3, 5, 7, and 10). SVM employed linear and RBF kernels, while Random Forest utilized parameter tuning including *n_estimators*, *max_depth*, and *min_samples_split*.

The results show that the SVM method with the RBF kernel achieved the best performance at K-Fold 10 with parameters C = 7 and gamma = 0.001, yielding an accuracy of 81.30%, precision of 82.21%, recall of 81.30%, and F1-score of 81.27%. Meanwhile, Random Forest performed best at K-Fold 7 with parameters *n_estimators* = 100, *max_depth* = 40, and *min_samples_split* = 2, producing an accuracy of 78.70%, precision of 78.98%, recall of 78.70%, and F1-score of 78.63%.

Keywords: Sentiment Analysis, NusaX, Dayak Jalai Language, SVM, Random Forest, K-Fold.