

ABSTRAK

Perkembangan *Natural Language Processing* (NLP) mendorong pemanfaatan analisis teks pada berbagai bidang, termasuk klasifikasi teks. Namun, belum terdapat kesimpulan yang konsisten mengenai metode representasi teks yang memberikan performa terbaik pada klasifikasi kategori produk. Oleh karena itu, penelitian ini bertujuan membandingkan performa metode *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Word2Vec* sebagai representasi teks pada klasifikasi kategori produk Tokopedia menggunakan algoritma *Support Vector Machine* (SVM). Penelitian dilakukan melalui tahapan *preprocessing* teks, pembagian data menggunakan rasio 80:20, penanganan ketidakseimbangan data menggunakan *Synthetic Minority Oversampling Technique* (SMOTE), representasi teks, pelatihan model SVM, validasi menggunakan *Stratified 5-Fold Cross Validation*, serta evaluasi menggunakan *accuracy*, *macro precision*, *macro recall*, dan *macro F1-score*. Hasil penelitian menunjukkan bahwa TF-IDF memberikan performa lebih baik dibandingkan *Word2Vec* dengan *accuracy* sebesar 95,91%, *macro precision* 92,90%, *macro recall* 92,83%, dan *macro F1-score* 92,84%, sedangkan *Word2Vec* memperoleh *accuracy* 89,83%, *macro precision* 83,16%, *macro recall* 85,47%, dan *macro F1-score* 84,17%. Hasil tersebut menunjukkan bahwa TF-IDF lebih efektif dibandingkan *Word2Vec* dalam klasifikasi kategori produk pada dataset ulasan Tokopedia yang didominasi ulasan singkat dengan rata-rata panjang 8,3 kata.

Kata kunci: Klasifikasi Teks, TF-IDF, *Word2Vec*, SMOTE, SVM

ABSTRACT

The advancement of Natural Language Processing (NLP) has encouraged the application of text analysis in various domains, including text classification. However, there is still no consistent conclusion regarding the text representation method that provides the best performance for product category classification. Therefore, this study aims to compare the performance of Term Frequency–Inverse Document Frequency (TF-IDF) and Word2Vec as text representation methods for Tokopedia product category classification using the Support Vector Machine (SVM) algorithm. The research was conducted through text preprocessing, data splitting with an 80:20 ratio, handling class imbalance using the Synthetic Minority Oversampling Technique (SMOTE), text representation, SVM model training, validation using Stratified 5-Fold Cross Validation, and evaluation based on accuracy, macro precision, macro recall, and macro F1-score. The results show that TF-IDF outperformed Word2Vec, achieving an accuracy of 95.91%, a macro precision of 92.90%, a macro recall of 92.83%, and a macro F1-score of 92.84%. In comparison, Word2Vec achieved an accuracy of 89.83%, a macro precision of 83.16%, a macro recall of 85.47%, and a macro F1-score of 84.17%. These findings indicate that TF-IDF is more effective than Word2Vec for product category classification on the Tokopedia review dataset, which is dominated by short reviews with an average length of 8.3 words.

Keywords: Text Classification, TF-IDF, Word2Vec, SMOTE, SVM