ABSTRAK

Klasifikasi sentimen merupakan salah satu teknik dalam pengolahan teks yang digunakan untuk mengidentifikasi opini pengguna terhadap suatu produk atau layanan. Salah satu permasalahan utama dalam klasifikasi sentimen adalah keterbatasan data berlabel yang berkualitas, yang dapat mempengaruhi performa model. Proses pelabelan data secara manual membutuhkan waktu dan biaya yang cukup besar, terutama pada dataset dengan jumlah data yang besar. Oleh karena itu, diperlukan pendekatan yang mampu memanfaatkan data tidak berlabel secara efektif untuk meningkatkan performa model tanpa harus melakukan pelabelan secara keseluruhan. Penelitian ini bertujuan untuk menganalisis penerapan metode *semi-supervised learning* berbasis *pseudo-labeling* dalam meningkatkan performa klasifikasi sentimen pada data ulasan pengguna.

Metode yang digunakan meliputi pengumpulan data, *preprocessing*, vektorisasi teks menggunakan Count Vectorizer dan TF-IDF dengan pendekatan *n-gram*, pelatihan model menggunakan Naive Bayes dan Random Forest, serta penerapan *pseudo-labeling* secara iteratif dengan variasi nilai *threshold* dari 0,6 hingga 0,9. Dataset yang digunakan terdiri dari dataset IndoNLU, E-commerce, dan ulasan aplikasi BCA Mobile. Evaluasi model dilakukan menggunakan metrik *accuracy* dan *F1-score* untuk membandingkan performa sebelum dan setelah penerapan *semi-supervised learning*.

Hasil penelitian menunjukkan bahwa penerapan *semi-supervised learning* mampu meningkatkan performa model pada sebagian besar skenario pengujian. Pada dataset E-commerce, akurasi Naive Bayes dengan TF-IDF meningkat dari 0,913 menjadi 0,932 pada threshold 0,9, sedangkan Random Forest meningkat dari 0,893 menjadi 0,917 pada threshold 0,6. Pada dataset IndoNLU, peningkatan akurasi tertinggi terjadi pada Random Forest dengan TF-IDF dari 0,827 menjadi 0,838 pada threshold 0,8. Sementara itu, pada dataset BCA Mobile, peningkatan performa relatif kecil dan cenderung fluktuatif, dengan beberapa skenario mengalami penurunan performa. Secara keseluruhan, rata-rata akurasi meningkat dari 0,895 menjadi 0,900 pada Naive Bayes berbasis Count Vectorizer, dan dari 0,887 menjadi 0,897 pada Random Forest berbasis TF-IDF. Hasil ini menunjukkan bahwa Random Forest memiliki performa yang lebih stabil dalam memanfaatkan data tambahan, sementara kualitas pseudo-label sangat dipengaruhi oleh nilai *threshold* yang digunakan.

Kata Kunci: Klasifikasi Sentimen, *Semi-Supervised Learning*, *Pseudo-Labeling*, Naive Bayes, Random Forest

ABSTRACT

Sentiment classification is a text processing technique used to identify user opinions toward a product or service. One of the main challenges in sentiment classification is the limited availability of high-quality labeled data, which can affect model performance. Manual data labeling requires significant time and cost, especially when dealing with large datasets. Therefore, an approach that can effectively utilize unlabeled data is needed to improve model performance without relying entirely on manual annotation. This study aims to analyze the application of a semi-supervised learning method based on pseudo-labeling to enhance sentiment classification performance on user review data.

The proposed method consists of data collection, preprocessing, text vectorization using Count Vectorizer and TF-IDF with an n-gram approach, model training using Naive Bayes and Random Forest, and iterative pseudo-labeling with threshold values ranging from 0.6 to 0.9. The datasets used include IndoNLU, E-commerce, and BCA Mobile user reviews collected from the Google Play Store. Model evaluation is conducted using accuracy and F1-score metrics to compare performance before and after applying semi-supervised learning.

The experimental results show that semi-supervised learning improves model performance in most scenarios. On the E-commerce dataset, the accuracy of Naive Bayes with TF-IDF increased from 0.913 to 0.932 at a threshold of 0.9, while Random Forest improved from 0.893 to 0.917 at a threshold of 0.6. On the IndoNLU dataset, the highest improvement was observed in Random Forest with TF-IDF, increasing from 0.827 to 0.838 at a threshold of 0.8. Meanwhile, on the BCA Mobile dataset, performance improvements were relatively small and fluctuated, with some scenarios showing slight decreases. Overall, the average accuracy increased from 0.895 to 0.900 for Naive Bayes with Count Vectorizer, and from 0.887 to 0.897 for Random Forest with TF-IDF. These results indicate that Random Forest provides more stable performance in utilizing pseudo-labeled data, while the effectiveness of pseudo-labeling is highly dependent on the selected threshold.

Keywords: *Sentiment Classification, Semi-Supervised Learning, Pseudo-Labeling, Naive Bayes, Random Forest*