

ABSTRAK

Pariwisata pantai di Kabupaten Pamekasan memiliki potensi yang besar dalam mendukung pertumbuhan ekonomi daerah, namun pemanfaatan ulasan wisatawan sebagai sumber informasi untuk evaluasi kualitas destinasi masih belum dilakukan secara optimal. Ulasan yang tersedia pada berbagai *platform* digital mengandung informasi penting mengenai pengalaman dan persepsi pengunjung, tetapi karakteristik teks yang tidak terstruktur, adanya kesalahan penulisan, serta ketidakseimbangan distribusi kelas menyebabkan proses analisis sentimen menjadi kurang efektif. Penelitian ini bertujuan untuk mengembangkan model analisis sentimen yang mampu meningkatkan kualitas klasifikasi ulasan wisata pantai melalui integrasi metode *Damerau Levenshtein Distance* (DLD), *Text Augmentation*, *Term Frequency-Inverse Document Frequency* (TF-IDF), dan *Multinomial Naive Bayes* (MNB). Penelitian ini menggunakan sebanyak 2.473 ulasan yang dikumpulkan dari Google Maps, TripAdvisor, TikTok, YouTube, dan Facebook. Data terlebih dahulu diproses melalui tahapan *preprocessing*. Selanjutnya dilakukan koreksi ejaan menggunakan DLD, pelabelan sentimen berbasis leksikon INSET, serta augmentasi data untuk mengurangi ketidakseimbangan kelas sehingga jumlah data meningkat menjadi 2.752 ulasan. Fitur teks kemudian diekstraksi menggunakan TF-IDF dan diklasifikasikan menggunakan *Multinomial Naive Bayes*. Untuk memperoleh konfigurasi model terbaik, dilakukan optimasi parameter menggunakan *GridSearchCV*, kemudian hasilnya dibandingkan dengan algoritma *Support Vector Machine* (SVM). Hasil penelitian menunjukkan bahwa konfigurasi terbaik *Multinomial Naive Bayes* dengan parameter $\alpha = 0,01$ menghasilkan akurasi sebesar 89%, lebih tinggi dibandingkan *Support Vector Machine* yang memperoleh akurasi sebesar 81,54%. Hasil penelitian ini menunjukkan bahwa integrasi koreksi ejaan, augmentasi data, dan representasi fitur TF-IDF mampu meningkatkan kualitas representasi teks serta menghasilkan model klasifikasi sentimen yang akurat, stabil, dan mudah diinterpretasikan. Model yang dikembangkan dapat menjadi pendukung pengambilan keputusan berbasis data dalam pengelolaan dan pengembangan destinasi wisata pantai di Kabupaten Pamekasan serta menjadi referensi bagi penelitian analisis sentimen berbahasa Indonesia pada bidang pariwisata.

Kata kunci: *Damerau Levenshtein Distance*, Inovasi dan Infrastruktur, *Naive Bayes Multinomial*, Analisis Sentimen, dan TF-IDF.

ABSTRACT

Coastal tourism in Pamekasan Regency holds significant potential for driving regional economic growth; however, the utilization of tourist reviews to evaluate destination quality remains suboptimal. While reviews across digital platforms contain valuable insights into visitor experiences and perceptions, challenges such as unstructured text, spelling errors, and class imbalance hinder the effectiveness of sentiment analysis. This study aims to develop a sentiment analysis model that improves the classification quality of coastal tourism reviews by integrating Damerau-Levenshtein Distance (DLD), Text Augmentation, Term Frequency-Inverse Document Frequency (TF-IDF), and Multinomial Naive Bayes (MNB). The study utilized a dataset of 2,473 reviews collected from Google Maps, TripAdvisor, TikTok, YouTube, and Facebook. Following initial preprocessing, spelling correction was performed using DLD, sentiment labeling was conducted based on the INSET lexicon, and data augmentation was applied to mitigate class imbalance, increasing the total dataset to 2,752 reviews. Text features were then extracted using TF-IDF and classified using Multinomial Naive Bayes. To identify the optimal model configuration, parameter optimization was performed using GridSearchCV, and the results were compared against the Support Vector Machine (SVM) algorithm. The findings indicate that the optimal Multinomial Naive Bayes configuration (with $\alpha = 0.01$) achieved an accuracy of 89.63%, outperforming the Support Vector Machine, which achieved 81.54%. The results demonstrate that integrating spelling correction, data augmentation, and TF-IDF feature representation enhances text representation quality and yields a sentiment classification model that is accurate, stable, and interpretable. The developed model can support data-driven decision-making for the management and development of coastal tourism destinations in Pamekasan Regency and serve as a reference for future research on Indonesian-language sentiment analysis in the tourism sector.

Keywords: Damerau-Levenshtein Distance, Innovation and Infrastructure, Multinomial Naive Bayes, Sentiment Analysis, and TF-IDF.