

ABSTRAK

Ulasan daring mengenai destinasi wisata pantai di Kabupaten Sumenep merupakan sumber data krusial, namun analisisnya terhambat oleh data yang tidak terstruktur dan penuh *noise*, terutama kesalahan ketik (*typo*), penggabungan kata yang tidak tepat, serta kendala efisiensi dalam pelabelan data manual. Penelitian ini bertujuan untuk mengatasi tantangan tersebut dengan membangun model analisis sentimen multi-aspek menggunakan *IndoBERT*. Penelitian ini mengintegrasikan alur pra-pemrosesan modul koreksi ejaan cerdas yang menggabungkan *Damerau-Levenshtein Distance* dengan *N-Gram*. Selain itu, telah diterapkan metode pelabelan semi-otomatis menggunakan *Zero-Shot Classification* berbasis XNLI untuk efisiensi proses anotasi. Teknik Teks Augmentasi (*Synonym Replacement* dan *Back-Translation*) juga digunakan untuk menangani ketidakseimbangan data pada kelas minoritas. Menggunakan kerangka kerja CRISP-DM, kinerja model dievaluasi melalui tiga skenario pengujian. Hasil penelitian menunjukkan bahwa model *Baseline* (Skenario 1) mencatatkan akurasi numerik tertinggi sebesar 96,12% dan *Macro F1-Score* 96,11%. Namun, Skenario 3 (Koreksi + Augmentasi) terbukti sebagai model yang paling tangguh secara strategis dengan akurasi 95,14%, karena terbukti mampu meningkatkan performa klasifikasi pada aspek pariwisata yang paling sulit diprediksi, yaitu aspek *Accessibility*. Meskipun terdapat penurunan akurasi sebesar 0,98% akibat pergeseran fitur kata kunci pada proses perbaikan ejaan, integrasi augmentasi teks berhasil membuat model lebih sensitif terhadap ulasan yang kompleks. Model tersebut telah diimplementasikan ke dalam aplikasi berbasis web menggunakan *Streamlit* untuk memberikan wawasan strategis bagi pengembangan pariwisata di Kabupaten Sumenep.

Kata Kunci: Analisis Sentimen Multi-Aspek, *IndoBERT*, *Damerau-Levenshtein Distance*, *N-Gram*, Teks Augmentasi, *Zero-Shot Classification*.

ABSTRACT

Online reviews regarding coastal tourism destinations in Sumenep Regency serve as crucial data sources; however, their analysis is hindered by unstructured and noisy data, particularly typos, improper word merging, and efficiency constraints in manual data labeling. This study aims to address these challenges by developing a multi-aspect sentiment analysis model using IndoBERT. This research integrates a preprocessing pipeline featuring an intelligent spelling correction module that combines Damerau-Levenshtein Distance with N-Gram. Furthermore, a semi-automatic labeling method using XNLI-based Zero-Shot Classification was applied to enhance the efficiency of the annotation process. Text Augmentation techniques (Synonym Replacement and Back-Translation) were also utilized to address data imbalance in minority classes. Using the CRISP-DM framework, the model's performance was evaluated through three testing scenarios. The results indicated that the Baseline model (Scenario 1) recorded the highest numeric accuracy of 96.12% and a Macro F1-Score of 96.11%. However, Scenario 3 (Correction + Augmentation) proved to be the most strategically robust model with an accuracy of 95.14%, as it demonstrated the ability to improve classification performance on the most difficult-to-predict tourism aspect, namely Accessibility. Although there was a 0.98% decrease in accuracy due to shifts in keyword features during the spelling correction process, the integration of text augmentation successfully rendered the model more sensitive to complex reviews. The model has been implemented into a web-based application using Streamlit to provide strategic insights for tourism development in Sumenep Regency.

Keywords: *Multi-Aspect Sentiment Analysis, IndoBERT, Damerau-Levenshtein Distance, N-Gram, Text Augmentation, Zero-Shot Classification.*