

ABSTRAK

Ulasan pada media sosial seperti Google Maps dan TikTok menjadi sumber informasi penting dalam menilai persepsi wisatawan terhadap suatu destinasi wisata, khususnya pada objek penelitian Air Terjun Tumpak Sewu. Namun, data ulasan umumnya bersifat tidak terstruktur, multibahasa, serta mengandung *slang*, emoji, dan *code-mixing*, sehingga sulit dianalisis menggunakan metode tradisional seperti TF-IDF dan *Naïve Bayes*. Penelitian ini mengusulkan penggunaan *Sentence*-BERT (SBERT) untuk ekstraksi fitur berbasis *embedding* semantik yang dikombinasikan dengan algoritma XGBoost sebagai model klasifikasi. Selain itu, dilakukan perbandingan teknik penyeimbangan data, yaitu SMOTE (*oversampling*) dan *Random Under Sampling* (RUS), serta penerapan proses *translation* untuk meningkatkan kualitas data. Proses penelitian meliputi pengambilan data, *preprocessing*, ekstraksi fitur, dan evaluasi model melalui empat skenario, yaitu *baseline*, *baseline* dengan SMOTE/RUS, *translate*, serta *translate* dengan SMOTE/RUS. Hasil penelitian menunjukkan bahwa kombinasi SBERT dan XGBoost mampu meningkatkan kualitas representasi teks dan performa model. Skenario 3 (*translate* tanpa *balancing*) menghasilkan performa terbaik pada data *testing* dengan akurasi 75,22% dan 86,53% pada data *training*. Sementara itu, skenario 4 (*translate* dengan RUS) menghasilkan akurasi *training* tertinggi sebesar 90,65% dan *testing* 73,51%, namun mengindikasikan potensi *overfitting*. Dengan demikian, *translation* berpengaruh paling signifikan, sedangkan *balancing* tidak selalu meningkatkan performa model. Pendekatan ini berpotensi diterapkan pada berbagai domain analisis sentimen multibahasa berbasis media sosial.

Kata kunci: Analisis Sentimen Multibahasa, *Sentence*-BERT, SMOTE, RUS, XGBoost, TikTok, Google Maps, Tumpak Sewu.

ABSTRACT

Reviews on social media platforms such as Google Maps and TikTok have become an important source of information for assessing tourists' perceptions of a tourist destination, particularly in this study of Tumpak Sewu Waterfall. However, the review data are generally unstructured, multilingual, and contain slang, emojis, and code-mixing, making them difficult to analyze using traditional methods such as TF-IDF and Naïve Bayes. This study proposes the use of Sentence-BERT (SBERT) for semantic embedding-based feature extraction combined with the XGBoost algorithm as a classification model. In addition, a comparison of data balancing techniques, namely SMOTE (oversampling) and Random Under Sampling (RUS), as well as a translation process, is conducted to improve data quality. The research process includes data collection, preprocessing, feature extraction, and model evaluation through four scenarios: baseline, baseline with SMOTE/RUS, translated data, and translated data with SMOTE/RUS. The results show that the combination of SBERT and XGBoost improves text representation quality and overall model performance. Scenario 3 (translation without balancing) achieved the best performance on testing data with an accuracy of 75.22% and 86.53% on training data. Meanwhile, Scenario 4 (translation with RUS) produced the highest training accuracy of 90.65% and testing accuracy of 73.51%, but indicated potential overfitting. Thus, translation has the most significant impact on improving model performance, while data balancing does not always lead to performance improvement. This approach has the potential to be applied in various domains of multilingual sentiment analysis based on social media.

Keywords: *Multilingual Sentiment Analysis, Sentence-BERT, SMOTE, RUS, XGBoost, TikTok, Google Maps, Tumpak Sewu.*