

ABSTRAK

Perhatian publik terhadap perlindungan data pribadi terus meningkat seiring pesatnya perkembangan teknologi digital, menjadikannya isu penting untuk dianalisis. Opini publik di media sosial menjadi sumber informasi potensial untuk memahami persepsi masyarakat. Penelitian ini bertujuan untuk mengevaluasi performa algoritma Logistic Regression dalam mengklasifikasikan sentimen publik terkait isu perlindungan data, dengan membandingkan dua teknik representasi fitur, yaitu TF-IDF dan Word2Vec, serta penerapan SMOTE untuk mengatasi ketidakseimbangan distribusi data. Data dikumpulkan dari media sosial X (Twitter) sebanyak 1.191 ulasan dan diseleksi menjadi 1.159 data setelah penghapusan duplikat. Tahapan preprocessing meliputi tokenisasi, penghapusan *stopword*, dan *stemming*. Hasil pengujian menunjukkan bahwa kombinasi TF-IDF dengan Logistic Regression memberikan akurasi tertinggi sebesar 84%, sedangkan kombinasi Word2Vec dengan Logistic Regression menghasilkan akurasi sebesar 83%. Penerapan SMOTE juga terbukti meningkatkan performa model dengan membantu menyeimbangkan klasifikasi antarkelas. Temuan ini menunjukkan bahwa pemilihan teknik representasi fitur dan penanganan data tidak seimbang merupakan faktor krusial dalam meningkatkan kualitas analisis sentimen publik terkait perlindungan data.

Kata Kunci: Perlindungan Data, Sentimen, TF-IDF, Word2Vec, SMOTE, Logistic Regression.

ABSTRACT

Public attention to personal data protection continues to increase in line with the rapid development of digital technology, making it an important issue to analyze. Public opinion on social media is a potential source of information for understanding public perceptions. This study aims to evaluate the performance of the Logistic Regression algorithm in classifying public sentiment related to data protection issues, by comparing two feature representation techniques, namely TF-IDF and Word2Vec, as well as the application of SMOTE to address data distribution imbalance. Data was collected from social media platform X (Twitter) comprising 1,191 reviews, which were filtered down to 1,159 data points after duplicate removal. The preprocessing stages included tokenization, stopword removal, and stemming. The test results showed that the combination of TF-IDF with Logistic Regression provided the highest accuracy of 84%, while the combination of Word2Vec with Logistic Regression resulted in an accuracy of 83%. The application of SMOTE was also found to improve model performance by helping to balance classification across classes. These findings indicate that the selection of feature representation techniques and handling of imbalanced data are crucial factors in improving the quality of public sentiment analysis related to data protection.

Keywords: Data Protection, Sentiment, TF-IDF, Word2Vec, SMOTE, Logistic Regression.