ABSTRAK

Wabah penyakit campak di Indonesia, termasuk Kabupaten Sumenep, menimbulkan beragam opini masyarakat melalui media sosial X (Twitter). Analisis sentimen digunakan untuk mengidentifikasi kecenderungan sentimen pada tweet terkait wabah campak. Penelitian ini bertujuan menganalisis sentimen dan juga membandingkan kinerja ekstraksi fitur *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW) menggunakan algoritma *Logistic Regression*, serta mengetahui pengaruh penerapan *Synthetic Minority Over-sampling Technique* (SMOTE) terhadap kinerja klasifikasi. Data diperoleh melalui *crawling* pada platform X menggunakan kata kunci terkait wabah campak, dengan 6.144 tweet yang dikumpulkan pada periode Oktober 2024 hingga Oktober 2025. Rangkaian penelitian meliputi pengumpulan data, *preprocessing* teks, pelabelan sentimen, pembagian data, ekstraksi fitur TF-IDF dan BoW, klasifikasi *Logistic Regression*, serta evaluasi menggunakan *accuracy, precision, recall,* dan *F1-score*. Hasil pengujian tanpa SMOTE menghasilkan TF-IDF memperoleh *accuracy* 76,8%, *weighted precision* 76,9%, *weighted recall* 76,8%, dan *weighted F1-score* 76,8%, sedangkan BoW memperoleh *accuracy* 74,8%, *weighted precision* 75,1%, *weighted recall* 74,8%, dan *weighted F1-score* 74,9%. TF-IDF unggul 2,0 poin persentase berdasarkan *accuracy*. Setelah SMOTE, TF-IDF dan BoW memperoleh *accuracy* yang sama, yaitu 76,6%. Penerapan SMOTE meningkatkan *accuracy* BoW sebesar 1,8 poin persentase, sedangkan TF-IDF mengalami penurunan 0,2 poin persentase. Dengan demikian, TF-IDF dengan *Logistic Regression* tanpa SMOTE menghasilkan kinerja tertinggi dengan *accuracy* 76,8%.

Kata kunci: Analisis Sentimen; Wabah Campak; TF-IDF; Bag of Words; Logistic Regression; X/Twitter.

ABSTRACT

The measles outbreak in Indonesia, including Sumenep Regency, has sparked diverse public opinions on the social media platform X (Twitter). Sentiment analysis was employed to identify sentiment trends in tweets related to the measles outbreak. This study aimed to analyze sentiment, compare the performance of Term Frequency–Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) feature extraction methods using the Logistic Regression algorithm, and determine the impact of applying the Synthetic Minority Over-sampling Technique (SMOTE) on classification performance. Data were obtained by crawling the X platform using keywords related to the measles outbreak, resulting in a dataset of 6,144 tweets collected between October 2024 and October 2025. The research process encompassed data collection, text preprocessing, sentiment labeling, data splitting, TF-IDF and BoW feature extraction, Logistic Regression classification, and performance evaluation using accuracy, precision, recall, and F1-score metrics. Testing results without SMOTE showed that TF-IDF achieved an accuracy of 76.8%, weighted precision of 76.9%, weighted recall of 76.8%, and weighted F1-score of 76.8%, whereas BoW achieved an accuracy of 74.8%, weighted precision of 75.1%, weighted recall of 74.8%, and weighted F1-score of 74.9%. TF-IDF outperformed BoW by 2.0 percentage points in terms of accuracy. Following the application of SMOTE, both TF-IDF and BoW achieved an identical accuracy of 76.6%. The application of SMOTE increased BoW's accuracy by 1.8 percentage points, while TF-IDF experienced a decrease of 0.2 percentage points. Consequently, the combination of TF-IDF and Logistic Regression without SMOTE yielded the highest performance, achieving an accuracy of 76.8%.

Keywords: Sentiment Analysis; Measles Outbreak; TF-IDF; Bag of Words; Logistic Regression; X/Twitter.