Evaluating Preprocessing Strategies and Twitter Sentiment Analysis for the 2024 Simultaneous Regional Elections

Asro[1], Eka Purnama Harahap[2], Patah Herwanto[3], Agus Sulaiman[4]
[1,3]Stmik im, Indonesia
[2,4]Universitas Raharja, Indonesia
Corresponding Author : asroharun6@gmail.com

Abstract

The 2024 Simultaneous Regional Elections (Pilkada) in Indonesia have garnered significant attention on social media platforms, especially Twitter. This digital space serves as a vital arena for citizens to voice their opinions about election candidates, policies, and the overall electoral process. This research aims to evaluate the impact of various data preprocessing strategies on sentiment analysis regarding the 2024 Regional Election. The implemented preprocessing steps include text normalization, stop word removal, tokenization, and stemming. These techniques are pivotal in enhancing the accuracy of sentiment classification. In this study, two vectorization methods—TF-IDF and Count Vectors—were employed to assess the effectiveness of Random Forest classification models in distinguishing between different sentiments expressed. The findings reveal that preprocessing markedly improves model accuracy, with TF-IDF achieving a 79.02% accuracy rate and Count Vectors attaining a slightly higher rate of 79.72%. The sentiment analysis identified that the majority of the detected sentiments were positive (67.7%), followed by negative (26.4%), and neutral (5.9%). This analysis underscores the essential role of data preprocessing in refining sentiment analysis and provides insightful perspectives on public sentiments towards the 2024 Regional Election in Indonesia.

Keywords: 2024 Regional Elections, Sentiment Analysis, Data Preprocessing, Twitter, Naive Bayes, Random Forest.