ABSTRACT

IMPLEMENTATION OFNORMALIZATION TEXT FOR SENTIMENT ANALYSIS OF INDONESIAN TEXTS USING MACHINE LEARNING

by

Indah Purwanti

20110007

Sentiment analysis is a major focus in natural language processing (NLP) aimed at recognizing feelings or sentiments based on opinions in text. Datasets from online media often contain slang words and misspellings, which pose challenges in text sentiment classification due to their non-standard forms and significant noise. This noise can reduce the accuracy of sentiment analysis models because the models struggle to recognize text.Therefore, text normalization becomes an essential step in preprocessing to reduce noise, ensure data consistency, and improve model performance. This study aims to analyse the impact of text normalization at the preprocessing stage on the accuracy of Indonesian sentiment analysis models by creating a slang corpus from a combination of internet dictionaries and slang words in the dataset. The data used are product reviews from Lazada and Tokopedia on the Kaggle platform. The results show that the application of text normalization significantly increases the accuracy of sentiment analysis models. The Naïve Bayes model performs better than the Decision Tree model, with the accuracy of the Naïve Bayes model after normalization increasing from 88.03% to 88.79% for Lazada data and from 88.23% to 90.55% for Tokopedia data. However, the Decision Tree model on Tokopedia data experienced a slight decrease in accuracy from 88.32% to 88.12%. The application of the slang word corpus also improves recall for negative cases, although precision tends to decrease after normalization, indicating a trade-off. Nevertheless, the model remains effective in identifying positive sentiments after text normalization.

Keyword: *Sentiment classification, machine learning, text normalization, and Indonesian texts*