Sentiment Analysis and Topic Modelling on English-Language Reviews of the Novel *Yellowface* Using Support Vector Machine (SVM) and Latent Dirichlet Allocation (LDA)

Nur Zuzzaifa
Data Science Study Program, Faculty of Science and Technology
University of Technology Yogyakarta
Jl. Ringroad Utara Jombor Sleman Yogyakarta
E-mail: nurzuzzaifa@gmail.com

ABSTRACT

This study aims to classify readers' reviews of Yellowface by R.F. Kuang into positive and negative sentiments and to identify the dominant topics within each sentiment group. The dataset consists of 10,000 English-language reviews collected from the Goodreads platform. The preprocessing stages include case folding, punctuation removal, word normalization, stopword removal, tokenization, and lemmatization. Feature extraction was performed using Term Frequency-Inverse Document Frequency (TF-IDF) for sentiment classification and Bag of Words for topic modelling. The classification model was built using the Support Vector Machine (SVM) algorithm, which demonstrated strong performance, achieving 95% accuracy on training data and 86% on testing data. Precision, recall, and F1-score metrics confirmed that the model is balanced and not biased toward any particular class. Topic modelling was conducted using Latent Dirichlet Allocation (LDA) separately for each sentiment group. The results showed that the positive sentiment model achieved the highest coherence score of 5.706 with an optimal number of 4 topics, while the negative sentiment model reached a coherence score of 5.886 with 8 optimal topics. The positive topics reflected the novel's core themes, racial issues, cancel culture, and favorable reader responses. In contrast, the negative topics included criticism of the publishing industry, writing style, plot development, and disappointing reading experiences.

Keywords: Classification, Latent Dirichlet Allocation (LDA), Topic Modelling, Support Vector Machine (SVM), *Yellowface*.