Identifying Difficult Quran Verses Using *Random Forest* and SVM

Satya Rayyis Baruna[1], Kemas Muslim Lhaksmana[2]
[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]rayyis@students.telkomuniversity.ac.id,
[2]kemasmuslim@telkomuniversity.ac.id

Abstract

The Qur'an contains thousands of verses with diverse Arabic linguistic structures. Some are easy to understand, while others are difficult, especially for readers without strong Arabic proficiency. This poses a challenge in learning and understanding the Qur'an, particularly for beginners. However, automatic systems capable of identifying difficult verses based on linguistic characteristics are still limited. This study proposes an automatic classification system to identify the difficulty level of verses using a dataset of 625 verses labeled as "easy" or "difficult" by respondents. The Random Forest algorithm is applied as the main classification method, with Support Vector Machine (SVM) used for comparison. The texts undergo preprocessing, feature transformation using TF-IDF, and class balancing with Random Undersampling (RUS) and SMOTE. The experimental results show that, after applying class balancing, the Random Forest model improved its performance in recognizing difficult verses, with recall increasing from 14.89% to 63.83%, F1-score from 22.22% to 51.28%, and overall accuracy reaching 63.69%. Although there was a slight decrease in accuracy, the model achieved a more balanced performance across both classes. This study contributes to the development of an automatic verse classification system that supports more effective and targeted Qur'anic learning.

Keywords: qur'an, machine learning, text classification, random forest, TF-IDF, difficult verses