Identifikasi 10 Bahasa Daerah Indonesia Menggunakan Pembelajaran Mesin

Azhar Baihaqi Nugraha[1], Ade Romadhony[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]azharnugraha@students.telkomuniversity.ac.id, [2]aderomadhony@telkomuniversity.ac.id

Abstract

Language Identification (LI) plays a crucial role in recognizing the diverse regional languages used in Indonesia, which vary in writing and pronunciation. It is an application of Natural Language Processing (NLP) where LI is commonly addressed through Text Classification (TC) approaches. In this study, we conduct language identification on 10 Indonesian regional languages based on the NusaX dataset. The objective of LI is to determine the language used in a given context. We employ six methods, namely Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), Decision Tree (DT), Rocchio Classification (RC), Logistic Regression (LR), and Random Forest (RF), using two feature extraction techniques: N-gram and TF-IDF. The main goal of this research is to construct models for identifying regional languages and to evaluate the performance of these models using the two feature extraction methods. The results of our experiments demonstrate that the identification of Indonesian regional languages using these six models and two feature extraction methods achieves excellent performance. Notably, the NBC model exhibits the highest accuracy of 0.992 for TF-IDF and 0.994 for N-Gram. Error Analysis (EA) is performed on the test results to investigate the reasons for misclassification. EA reveals that misclassifications occur due to the presence of similar words in other languages and their dominant usage in those languages.

Keywords: ***teks classification, language identification, supervise learning machine, decision tree, naïve bayes classifier.***