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

Abstrak

Bahasa merupakan alat komunikasi yang digunakan oleh manusia untuk bersosialisasi. Namun Indonesia memiliki banyak bahasa daerah yang beragam cara penulisan dan penyebutannya, disinilah pengidentifikasian bahasa berperan. *Language Identification* (LI) merupakan salah satu pengaplikasian menggunakan *Natural Languange Processing* (NLP). LI umumnya diselesaikan menggunakan pendekatan *Text Classification* (TC), dimana pada tugas akhir ini akan dilakukan identifikasi terhadap 10 bahasa daerah Indonesia berdasarkan dataset NusaX. Tujuan LI adalah untuk mengetahui bahasa apa yang digunakan dalam suatu konteks. Metode yang digunakan untuk menyelesaikan task LI pada Tugas Akhir ini adalah *Support vector machine* (SVM), *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), *Rocchio Classification* (RC), *Logistic Regression* (LR), *Random Forest* (RF), dengan dua fitur yaitu N-gram dan TF-IDF. Tujuan dari penelitian ini adalah membangun model identifikasi bahasa daerah dan mengevaluasi kinerja dari enam metode dan dua fitur ekstraksi yang digunakan dalam melakukan pengidentifikasian 10 bahasa daerah Indonesia. Hasil pengujian menunjukkan bahwa identifikasi bahasa daerah Indonesia menggunakan enam model dan dua fitur menghasilkan performa yang sangat baik dengan model paling baik adalah NBC dengan akurasi 0.992 untuk TF-IDF dan 0.994 untuk N-Gram. *Error Analysis* (EA) dilakukan kepada hasil pengujian untuk mengetahui mengapa model dapat melakukan salah prediksi bahasa. EA menunjukkan penyebab salah prediksi bahasa adalah terdapat kata-kata yang mirip dalam bahasa lain dan mempunyai penyebaran kata yang lebih dominan pada bahasa lain.

Kata kunci : teks klasifikasi, identifikasi bahasa, *supervise learning machine, decision tree, naïve bayes classifier.*