ANALISIS PERBANDINGAN PERFORMA ALGORITMA *RANDOM FOREST* DAN *LOGISTIC REGRESSION* UNTUK DETEKSI BERITA HOAKS

Retno Anjani[1], Deni Arifianto[2], Ulya Anisatur Rosyidah[3]
[123]Program Studi Sistem Informasi, Fakultas Teknik,
Universitas Muhammadiyah Jember
Email: anjaniretno1125@gmail.com [1], deniarifianto@unmuhjember.ac.id [2],
ulyaanisatur@unmuhjember.ac.id[3]

ABSTRAK

Penyebaran informasi palsu atau hoaks di media digital Indonesia menjadi permasalahan serius yang memicu disinformasi publik. Namun, pengolahan data teks berita dunia nyata sering kali terkendala oleh masalah ketidakseimbangan distribusi kelas (*imbalanced data*), di mana jumlah berita hoaks jauh lebih mendominasi dibandingkan berita valid. Penelitian ini bertujuan untuk menganalisis dan membandingkan performa algoritma *Random Forest* dan *Logistic Regression* dalam mendeteksi berita hoaks berbahasa Indonesia, baik sebelum maupun sesudah penanganan *imbalanced data* menggunakan metode *Customized Instance Random Undersampling* (CIRUS). Dataset yang digunakan bersumber dari repositori *Kaggle* (Kominfo) sebanyak 17.185 artikel berita dengan 15.411 data bersih. Tahap ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) berkapasitas 10.000 fitur dengan gabungan atribut *tittle* dan *body_text*. Metode CIRUS diterapkan melalui pembentukan 32 *compact set* berbasis *K-Means* hingga diperoleh 10.274 data seimbang. Pengujian dilakukan menggunakan skema *K-Fold Cross Validation* (K=5) berbasis metrik *Accuracy, Precision*, *Recall*, *F1-Score*, dan *Confusion Matrix*. Hasil pengujian menunjukkan bahwa pada kondisi data tidak seimbang (sebelum CIRUS), *Random Forest* memperoleh akurasi 99,81% berbanding 99,51% pada *Logistic Regression*, namun nilai tersebut bersifat akurasi semu akibat bias kelas mayoritas. Setelah penanganan *imbalanced data* menggunakan CIRUS (data seimbang), *Random Forest* kembali terbukti mengungguli *Logistic Regression* secara signifikan dengan akurasi sebesar 95,16% berbanding 92,21%. Penelitian ini menyimpulkan bahwa penyeimbangan CIRUS efektif mengeliminasi bias kelas, di mana keunggulan *Random Forest* didorong oleh arsitektur *ensemble learning* dengan 100 *decision trees* yang mampu memetakan hubungan *non-linier* dan dimensi fitur TF-IDF yang tinggi secara lebih efektif dibanding pembatas linier pada *Logistic Regression.*

Kata kunci: Berita Hoaks, CIRUS, *Text Classification, Logistic Regression, Random Forest,* TF-IDF.