

ABSTRAK

Penyebaran berita *hoax* di Indonesia menjadi tantangan seiring perkembangan teknologi informasi dan media digital. Kemiripan karakteristik berita *hoax* dan berita asli menyebabkan identifikasi manual kurang efektif sehingga diperlukan sistem deteksi otomatis. Penelitian ini membandingkan performa *Long Short-Term Memory (LSTM)*, *Random Forest (RF)*, dan *Hybrid LSTM-Random Forest* dalam mendeteksi berita *hoax* berbahasa Indonesia. *Dataset* terdiri atas 4.502 berita yang seimbang antara berita *hoax* dari *TurnBackHoax* dan berita aktual dari Kompas, CNN Indonesia, serta Tempo. Tahapan penelitian meliputi *preprocessing* (*case folding*, pembersihan tanda baca dan simbol, tokenisasi, *stemming*, dan *stopword removal*), representasi fitur menggunakan TF-IDF pada *Random Forest* serta *word embedding* pada LSTM dan *hybrid*, pembagian data menjadi *training*, *validation*, dan *testing*, pelatihan model, serta evaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Metode *hybrid* memanfaatkan LSTM sebagai *feature extractor* dan *Random Forest* sebagai *classifier*. Hasil penelitian menunjukkan bahwa *Random Forest* dengan nilai *n_estimator* 200, dan *max_depth* 40 memberikan performa terbaik dengan *accuracy* 0.9660, *precision* 0.9730, *recall* 0.9586, dan *F1-score* 0.9657. Berdasarkan hasil tersebut, metode *standalone*, khususnya *Random Forest*, memberikan performa terbaik pada *dataset* yang digunakan.

Kata kunci: Deteksi Berita Hoax, *Long Short-Term Memory (LSTM)*, *Random Forest*, *Hybrid LSTM-Random Forest*, *Text Classification*, *Natural Language Processing (NLP)*.

ABSTRACT

The spread of fake news in Indonesia has become a challenge amid the development of information technology and digital media. The similarities between fake news and genuine news make manual identification less effective, necessitating an automated detection system. This study compares the performance of Long Short-Term Memory (LSTM), Random Forest (RF), and Hybrid LSTM-Random Forest in detecting Indonesian-language fake news. The dataset consists of 4,502 news articles, balanced between fake news from TurnBackHoax and real news from Kompas, CNN Indonesia, and Tempo. The research stages included preprocessing (case folding, punctuation and symbol cleaning, tokenization, stemming, and stopword removal); feature representation using TF-IDF for Random Forest and word embeddings for LSTM and the hybrid model; splitting the data into training, validation, and testing sets; model training; and evaluation using accuracy, precision, recall, and F1-score. The hybrid method utilizes LSTM as a feature extractor and Random Forest as a classifier. The results show that Random Forest with `n_estimator` set to 200 and `max_depth` set to 40 delivers the best performance, achieving an accuracy of 0.9660, precision of 0.9730, recall of 0.9586, and an F1-score of 0.9657. Based on these results, the standalone method—specifically Random Forest—delivers the best performance on the dataset used.

Keywords: Fake News Detection, Long Short-Term Memory (LSTM), Random Forest, Hybrid LSTM-Random Forest, Text Classification, Natural Language Processing (NLP).