INTISARI

Pengaruh Metode *Feature Extraction* Terhadap Kinerja *Random Forest* Dalam Klasifikasi Berita *Hoax* Berbahasa Indonesia

Oleh

Muhammad Rafi

16/394094/PA/17185

Hoax merupakan isu yang sangat problematik di Indonesia. Penyebaran *hoax* dianggap dapat mengakibatkan beragam konflik, perpecahan, dan pertikaian di beberapa bagian di Indonesia. Dalam mengantisipasi hal tersebut, beberapa penelitian terdahulu telah melakukan percobaan untuk melakukan klasifikasi hoax dalam bahasa Indonesia menggunakan pendekatan *supervised learning*, seperti SVM, random forest, dan deep learning, namun nilai akurasi yang dihasilkan kebanyakan masih di sekitar angka 70%.

Metode *feature extraction* dapat memberikan informasi tambahan dan mempermudah pembelajaran pada model *supervised learning* dengan mengubah data teks menjadi representasi vektor numerik. Penelitian ini menggunakan Word2Vec, TF-IDF, dan *bag-of-words* sebagai bentuk *feature extraction* untuk melakukan klasifikasi berita *hoax* berbahasa Indonesia dengan menggunakan model *random forest.* Pencarian *hyper-parameters* terbaik model *random forest* untuk setiap metode *feature extraction* dilakukan dengan menggunakan *randomized cross-validation search*.

Menggunakan dataset yang berisi 159 berita dengan label *valid* dan 91 berita dengan label *hoax*, hasil evaluasi dari 20% data *testing* menunjukan bahwa model *random forest* memiliki kinerja terbaik saat menggunakan fitur yang dihasilkan oleh Word2Vec skip-gram maupun CBOW dengan akurasi 90%, presisi 91%, *recall* 87%, dan *f1-score* 89%.

Kata kunci: Hoax, Klasifikasi Teks, Random Forest, Ekstraksi Fitur, Word Embedding

ABSTRACT

Effect Of Feature Extraction Methods In Random Forest Performance In Classifying Hoax News In Indonesian Language

By

Muhammad Rafi

16/394094/PA/17185

Hoax is a really problematic issue in Indonesia. The spreading of hoax may create conflicts and riots in some regions. To prevent this, researches of Indonesian hoax classification using supervised learning are done in the past using *supervised learning* models such as SVM, random forest, and deep learning, but the resulting accuracy is still around 70%.

Feature extraction method can give additional information. thus making it easier for classifier to learn the data by transforming textual data to numerical vector representation. This research uses Word2Vec, TF-IDF, dan bag-of-words to extract text features and random forest to classify Indonesian hoax news. Randomized cross-validation search is also used to find the best random forest hyper-parameters for each given feature extraction methods.

Using dataset consisting of 159 valid news and 91 hoax news, evaluation results on 20% testing data shows that random forest has the best performance when using features generated by either Word2Vec skip-gram or CBOW with accuracy of 90%, precision of 91%, recall of 87%, and *f1-score* of 89%.

Keywords: Hoax, Text Classification, Random Forest, Feature Extraction, Word Embedding