ABSTRAK

Perkembangan Perkembangan teknologi komunikasi digital mendorong meningkatnya penggunaan layanan pesan singkat (SMS) sebagai media penyampaian informasi. Namun, tingginya volume SMS juga diiringi dengan maraknya spam penipuan (*fraud*) yang merugikan pengguna, sehingga diperlukan metode otomatis untuk membedakan pesan asli (*ham*) dan spam penipuan secara akurat. Penelitian ini berfokus pada klasifikasi SMS berbahasa Indonesia dengan memanfaatkan pendekatan *deep learning* berbasis Bidirectional Encoder Representations from Transformers (BERT). Tujuan penelitian adalah menerapkan serta membandingkan kinerja tiga model pralatih, yaitu IndoBERT, IndoBERTweet, dan Multilingual BERT (mBERT) dalam mendeteksi spam penipuan pada SMS.

Dataset penelitian berupa SMS berbahasa Indonesia yang diproses melalui tahapan pra-pemrosesan teks dan tokenisasi menggunakan *tokenizer* masing-masing model. Model kemudian dilakukan *fine-tuning* untuk tugas klasifikasi biner (*ham* dan *fraud*). Evaluasi performa dilakukan menggunakan skema *Stratified 5-Fold Cross Validation* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, serta memberikan perhatian khusus pada *F1-score* kelas *fraud* karena kesalahan deteksi pada kelas ini berdampak lebih serius. Selain itu, dilakukan pengujian tambahan tanpa *cross validation* menggunakan skema pembagian data 70:15:15 sebagai pembanding, serta analisis kesalahan prediksi untuk mengidentifikasi karakteristik pesan yang masih sulit diklasifikasikan oleh model. Penelitian ini juga melakukan pengujian menggunakan dataset eksternal untuk melihat ketahanan (*robustness*) model terhadap perbedaan karakteristik data.

Hasil pengujian menunjukkan bahwa IndoBERT memberikan performa terbaik dan paling stabil pada evaluasi internal dengan akurasi 97,56% ± 0,79 dan *F1-score fraud* 96,73% ± 0,97, diikuti oleh IndoBERTweet dengan akurasi 97,11% ± 1,20 dan *F1-score fraud* 96,10% ± 1,62, serta mBERT dengan akurasi 95,78% ± 0,93 dan *F1-score fraud* 94,38% ± 1,35. Dari sisi efisiensi, IndoBERT juga menjadi model dengan waktu pelatihan rata-rata paling cepat dibanding dua model lainnya. Pada pengujian ketahanan (*robustness*) menggunakan dataset eksternal, IndoBERTweet menunjukkan performa rata-rata terbaik dengan akurasi 91,56% ± 5,11 dan *F1-score fraud* 90,46% ± 6,26. Temuan ini mengindikasikan bahwa pemilihan model pralatih berbahasa Indonesia seperti IndoBERT efektif untuk performa internal, sementara IndoBERTweet cenderung lebih tangguh pada variasi data di luar dataset pelatihan.

Kata kunci: Klasifikasi SMS, Spam Penipuan, BERT, IndoBERT, IndoBERTweet, mBERT, Fine-tuning, Cross Validation.

ABSTRACT

The development of digital communication technology has increased the use of Short Message Service (SMS) as a medium for delivering information. However, the high volume of SMS messages is also accompanied by the increasing spread of fraudulent spam messages, which can cause significant harm to users. Therefore, an automatic method is required to accurately distinguish legitimate messages (ham) from fraudulent spam (fraud). This study focuses on the classification of Indonesian-language SMS messages by utilizing a deep learning approach based on Bidirectional Encoder Representations from Transformers (BERT). The objective of this research is to implement and compare the performance of three pretrained models, namely IndoBERT, IndoBERTweet, and Multilingual BERT (mBERT), in detecting fraudulent spam SMS.

The dataset used in this research consists of Indonesian SMS messages that were processed through text preprocessing and tokenization using the tokenizer of each model. The models were then fine-tuned for a binary classification task (ham and fraud). Performance evaluation was conducted using the Stratified 5-Fold Cross Validation scheme with metrics including accuracy, precision, recall, and F1-score, with particular emphasis on the fraud class F1-score due to the higher risk associated with misclassification in this class. In addition, an extra evaluation without cross validation using a 70:15:15 data split was performed as a comparison, along with an error analysis to identify message characteristics that are difficult to classify. This study also performed external dataset testing to evaluate the robustness of the models against differences in data characteristics.

The results show that IndoBERT achieved the best and most stable performance in internal evaluation, with an accuracy of 97.56% ± 0.79 and a fraud F1-score of 96.73% ± 0.97, followed by IndoBERTweet with an accuracy of 97.11% ± 1.20 and a fraud F1-score of 96.10% ± 1.62, and mBERT with an accuracy of 95.78% ± 0.93 and a fraud F1-score of 94.38% ± 1.35. In terms of efficiency, IndoBERT also recorded the fastest average training time compared to the other two models. In the robustness evaluation using an external dataset, IndoBERTweet achieved the best average performance, with an accuracy of 91.56% ± 5.11 and a fraud F1-score of 90.46% ± 6.26. These findings indicate that Indonesian pretrained models such as IndoBERT are highly effective for internal performance, while IndoBERTweet tends to be more robust when dealing with data variations outside the training dataset.

Keywords: *SMS Classification, Fraudulent Spam, BERT, IndoBERT, IndoBERTweet, mBERT, Fine-tuning, Cross Validation.*