ABSTRACT

A key challenge in supervised spam detection for low-resource languages is the limited availability of labeled data. Producing large-scale annotated datasets requires substantial human effort, making the process both time-consuming and inefficient. To address this limitation, AI-assisted labeling has emerged as a viable alternative annotation process. This study explores an AI-assisted labeling approach using ChatGPT, further refined through ensemble learning based validation (EL-ChatGPT), to improve annotation quality and enhance model performance. Three labeling strategies were compared: human, ChatGPT, and EL-ChatGPT labeling. These approaches were applied to two Indonesian spam datasets and evaluated across three pretrained deep learning models: BERT Multilingual, IndoBERT, and XLM-RoBERTa. Results demonstrate that on the first dataset, IndoBERT achieved higher accuracy (95.96%) and F1-score (96.04%) with ChatGPT labels compared to human labels (94.95% accuracy, 95.17% F1-score). EL-ChatGPT further improved these results, yielding the lowest test loss and highest F1-score (96.12%). Performance gains were more pronounced on the second dataset, where EL-ChatGPT enhanced IndoBERT accuracy to 97.67% and F1-score to 97.75%, outperforming both other labeling methods. These findings highlight ChatGPT's potential as a scalable alternative for data annotation and demonstrate that EL-ChatGPT can effectively enhance the model outcomes, in low resource contexts.

Keywords— Spam Detection, Data Labeling, Deep Learning, ChatGPT, Ensemble Learning, BERT, Indonesian Dataset

INTI SARI

Tantangan utama dalam deteksi spam berbasis supervised untuk bahasa dengan sumber daya terbatas adalah keterbatasan data berlabel yang tersedia. Pembuatan dataset beranotasi dalam skala besar memerlukan upaya manusia yang substansial, sehingga prosesnya menjadi memakan waktu dan tidak efisien. Untuk mengatasi keterbatasan ini, pelabelan berbantuan AI telah muncul sebagai alternatif yang layak dalam proses anotasi. Studi ini mengeksplorasi pendekatan pelabelan berbantuan AI menggunakan ChatGPT, yang disempurnakan lebih lanjut melalui validasi berbasis ensemble learning (EL-ChatGPT), untuk meningkatkan kualitas anotasi dan kinerja model. Tiga strategi pelabelan dibandingkan: pelabelan manusia, ChatGPT, dan EL-ChatGPT. Ketiga pendekatan ini diterapkan pada dua dataset spam berbahasa Indonesia dan dievaluasi menggunakan tiga model deep learning pra-latih: BERT Multilingual, IndoBERT, dan XLM-RoBERTa. Hasil menunjukkan bahwa pada dataset pertama, IndoBERT mencapai akurasi lebih tinggi (95,96%) dan skor F1 (96,04%) dengan label dari ChatGPT dibandingkan dengan label manusia (94,95% akurasi, 95,17% skor F1). EL-ChatGPT semakin meningkatkan hasil tersebut, menghasilkan loss pengujian terendah dan skor F1 tertinggi (96,12%). Peningkatan performa lebih terlihat jelas pada dataset kedua, di mana EL-ChatGPT meningkatkan akurasi IndoBERT hingga 97,67% dan skor F1 hingga 97,75%, melampaui kedua metode pelabelan lainnya. Temuan ini menyoroti potensi ChatGPT sebagai alternatif yang dapat diskalakan untuk anotasi data dan menunjukkan bahwa EL-ChatGPT secara efektif dapat meningkatkan hasil model, khususnya dalam konteks bahasa dengan sumber daya terbatas.

Kata Kunci—deteksi spam, pelabelan data, *deep learning*, ChatGPT, *ensemble learning*, BERT, Indonesian dataset