

## INTISARI

### MODEL KLASIFIKASI BERBASIS *MULTICLASS CLASSIFICATION* DENGAN KOMBINASI INDOBERT *EMBEDDING* DAN *LONG SHORT-TERM MEMORY* UNTUK *TWEET* BERBAHASA INDONESIA

Oleh

THARIQ ISKANDAR ZULKARNAIN M P  
18/424198/PA/18303

Pencarian *tweet* pada aplikasi Twitter menggunakan kata kunci atau *hashtag* terkadang masih kurang akurat ketika menggunakan kata yang memiliki beberapa arti. Untuk itu diperlukan pengembangan model yang dapat mengategorikan *tweet* sesuai dengan konteksnya. Klasifikasi teks merupakan salah satu tugas pemrosesan bahasa alami yang dapat mengategorikan teks secara otomatis berdasarkan konteksnya dengan bantuan metode *machine learning* atau *deep learning*. Sudah terdapat beberapa penelitian tentang pengembangan model klasifikasi teks pada *dataset* berbahasa Indonesia, namun masih mungkin untuk ditingkatkan akurasi. Penelitian ini bertujuan untuk meningkatkan performa model klasifikasi teks dari penelitian sebelumnya, dengan menggabungkan model *pre-trained* IndoBERT dengan arsitektur *Long Short-Term Memory* (LSTM) dalam mengklasifikasikan *tweet* berbahasa Indonesia ke beberapa kategori. Model IndoBERT-LSTM dengan skenario kombinasi *hyperparameter* terbaik (*batch size* sebesar 16, *learning rate* sebesar  $2e-5$ , dan menggunakan *average pooling*) berhasil mendapatkan *F1-score* sebesar 98,90% pada *dataset* yang tidak termodifikasi (peningkatan 0,70% dari model Word2Vec-LSTM dan 0,40% dari model *fine-tuned* IndoBERT) dan 92,83% pada *dataset* yang telah termodifikasi (peningkatan 4,51% dari model Word2Vec-LSTM dan 0,69% dari model *fine-tuned* IndoBERT). Akan tetapi, peningkatan dari model *fine-tuned* IndoBERT tidak terlalu signifikan dan model Word2Vec-LSTM memiliki total waktu pelatihan yang jauh lebih cepat.

Kata-kata kunci: Klasifikasi Teks, *Tweet* Berbahasa Indonesia, IndoBERT, *Long Short-Term Memory*.

## ABSTRACT

### **CLASSIFICATION MODEL BASED ON MULTICLASS CLASSIFICATION WITH A COMBINATION OF INDOBERT EMBEDDING AND LONG SHORT-TERM MEMORY FOR INDONESIAN TWEET**

By

THARIQ ISKANDAR ZULKARNAIN M P  
18/424198/PA/18303

Searching tweets on the Twitter application using keywords or hashtags is sometimes still less accurate when using words that have multiple meanings. This requires the development of a model that can categorize tweets according to their context. Text classification is one of the natural language processing tasks that can automatically categorize text based on its context with the help of machine learning or deep learning methods. There have been several previous studies regarding the development of a text classification model on Indonesian language datasets, but it is still possible to improve its accuracy. This research aims to improve the performance of the text classification model from previous studies, by combining the IndoBERT pre-trained model with the Long Short-Term Memory (LSTM) architecture in classifying Indonesian-language tweets into several categories. The IndoBERT-LSTM model with the best hyperparameter combination scenario (batch size of 16, learning rate of  $2e-5$ , and using average pooling) managed to get an F1-score of 98.90% on the unmodified dataset (0.70% increase from the Word2Vec-LSTM model and 0.40% from the fine-tuned IndoBERT model) and 92.83% on the modified dataset (4.51% increase from the Word2Vec-LSTM model and 0.69% from the fine-tuned IndoBERT model). However, the improvement from the fine-tuned IndoBERT model is not very significant and the Word2Vec-LSTM model has a much faster total training time.

Keywords: Text Classification, Indonesian Tweets, IndoBERT, Long Short-Term Memory.