ABSTRACT

THE IMPACT OF NAMED ENTITY RECOGNITION AND DATA AUGMENTATION ON THE PERFORMANCE OF DEEP LEARNING MODELS IN TEXT CLASSIFICATION

By

Nur Annisa Putri Rezkia

This study aims to classify news texts from the InaCOVED dataset into event and non-event categories using MLP, CNN, and LSTM models. Synonym-based data augmentation through the Kateglo API was applied to improve class balance, while NER based on BERT and an LLM with a prompting approach was used to extract key entities (Person, Organization, Location, and Disease). Text representations were generated using Word2Vec and the Keras Embedding Layer and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The best performance was achieved by the CNN model with the Keras Embedding Layer and LLM-based NER using 100-dimensional vectors, obtaining an accuracy of 0.9597, precision of 0.9442, recall of 0.9772, F1-score of 0.9604, and ROC-AUC of 0.9888, indicating that entity extraction methods, vector representation, and data augmentation significantly influence classification performance.

Keywords: NER, LLM, BERT, Data Augmentation, Word Embedding, Text Classification.

ABSTRAK

PENGARUH *NAMED ENTITY RECOGNITION* DAN AUGMENTASI DATA TERHADAP KINERJA MODEL *DEEP LEARNING* DALAM KLASIFIKASI TEKS

Oleh

Nur Annisa Putri Rezkia

Penelitian ini bertujuan mengklasifikasikan teks berita pada dataset InaCOVED ke dalam kategori *event* dan *non-event* menggunakan model MLP, CNN, dan LSTM. Augmentasi data berbasis penggantian sinonim melalui API Kateglo diterapkan untuk meningkatkan keseimbangan kelas, sementara NER berbasis BERT dan LLM dengan pendekatan *prompting* digunakan untuk mengekstraksi entitas penting (*Person*, *Organization*, *Location*, dan *Disease*). Representasi teks dibentuk menggunakan Word2Vec dan Keras *Embedding* Layer, kemudian dievaluasi dengan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan ROC-AUC. Hasil terbaik diperoleh pada model CNN dengan Keras *Embedding* Layer dan NER berbasis LLM pada dimensi vektor 100 dengan *accuracy* 0,9597, *precision* 0,9442, *recall* 0,9772, *F1-score* 0,9604, dan ROC-AUC 0,9888, yang menunjukkan bahwa metode ekstraksi entitas, representasi vektor, dan augmentasi data berpengaruh signifikan terhadap kinerja klasifikasi.

Kata-kata kunci: NER, LLM, BERT, Augmentasi Data, *Word Embedding*, Klasifikasi Teks.