ABSTRAK

ANALISIS *WORD PREDICTION* DENGAN MENGGUNAKAN *LANGUAGE MODEL Bidirection Encoding Representations from Transfordmers* (BERT) PADA *DATASET* KALIMAT BAHASA INDONESIA

Oleh

AGHITA NAMIRA YULIZA

Bahasa Indonesia sebagai bahasa nasional memiliki peran penting dalam berbagai bidang, termasuk pengembangan teknologi pemrosesan bahasa alami (*Natural Language Processing*). Salah satu pendekatan modern dalam NLP adalah penggunaan model transformer-based seperti BERT (*Bidirectional Encodder Representations from Transformers*) untuk menyelesaikan tugas *Masked Langugae Modeling* (MLM), yaitu menebak token yang hilang dalam suatu kalimat berdasarkan konteksnya. Tujuan penelitian ini adalah untuk mengevaluasi kinerja model BERT pada kalimat bahasa Indonesia dengan dataset 27.600 baris kalimat bahasa Indonesia. Model dilatih dengan dua skema, yaitu tanpa augmentasi (skema 1) dan dengan teknik augmentasi data (skema 2). Hasil evaluasi menunjukkan bahwa skema 2 memberikan kinerja yang lebih baik, dengan akurasi sebesar 42,1% (top-1), 53,7% (top-3), dan 58,1% (top-5), dibandingkan dengan skema 1 yang menghasilkan akurasi 29% (top-1), 42,6% (top-3), dan 52,6% (top-5). Peningkatan ini menunjukkan bahwa penggunaan augmentasi data dapat meningkatkan variasi kalimat dalam pelatihan model, kemampuan prediktif model terhadap kata-kata yang dimasking dapat ditingkatkan.

Kata kunci: *BERT, Masked Language Modelling, Bahasa Indonesia,* NLP*;*

ABSTRACT

WORD PREDICTION ANALYSIS USING THE BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) LANGUAGE MODEL ON INDONESIAN SENTENCE DATASETS

By

AGHITA NAMIRA YULIZA

Indonesian, as the national language, plays a crucial role in various fields, including the development of Natural Language Processing (NLP) technologies. One modern approach in NLP is the use of transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) to perform Masked Language Modeling (MLM), which involves predicting missing tokens in a sentence based on context. This study aims to evaluate the performance of the BERT model on Indonesian sentences using a dataset of 27,600 Indonesian sentence entries. The model was trained using two schemes: without augmentation (Scheme 1) and with data augmentation techniques (Scheme 2). Evaluation results show that Scheme 2 provides better performance, with an accuracy of 42.1% (top-1), 53.7% (top-3), and 58.1% (top-5), compared to Scheme 1 which achieved an accuracy of 29% (top-1), 42.6% (top-3), and 52.6% (top-5). This improvement indicates that data augmentation can enhance the diversity of training sentences, thereby improving the model's predictive capability for masked words.

***Keywords**: BERT, Masked Language Modeling, Indonesian Language, NLP*