ABSTRACT

Name Entity Recognition (NER) is one of the method for extracting informaction such as location. NER can be solved using a machine learning (ML) approach. NER with good performance can produce more relevant location extraction than searching for words in the gazetterr dataset. Many studies related to English NER already have good machine learning model performance, but Indonesian NER still needs to be improved. This is because of the small number of NER datasets labeled Indonesian, which has an impact on the lack of the proposed method.

BERT is a machine learning algorithm that is trained on Natural Language Processing (NLP) Next Sentence Prediction (NSP) and Masked Language Modeling (MLM) tasks. However, for NER tasks it is necessary to do tuning first to get better performance. The fine-tuning process can be in the form of training with labeled datasets and adding layers when carrying out the training process with labeled datasets.

This study proposes fine-tuning BERT by combining CNN, Bi-LSTM, Bi-GRU, and CRF. The BERT used is a number of Indonesian and English Pre-Training models. The datasets in this study are NERGrit-IndoNLU, NERGrit-Corpus, NERUGM, and NERUI. From the predetermined scenarios, there are a total of 16 models with 15 ML combinations and 1 tuning fully connected layer. The test results show that some of the models proposed in this study produce the best results for each dataset. F1 micro dataset NERUI 95%, F1 micro NERUGM 84%, and F1 micro NERGRIT-Corpus 84%, and F1 macro dataset NERGRIT-IndoNLU 81%. However, the average NER on the BERT_CNN model produces the best accuracy of 95.43%, F1 micro 80.9%, and F1 macro 76.16%. The performance of NER on the model used is good enough so that location extraction can be used which is more relevant when compared to finding a location directly with a gazetteer.

Keywords: *Extraction Location, NER, BERT, Pre-Training, Machine Learning*

INTISARI

Name Entity Recognition (NER) merupakan salah satu metode untuk melakukan ekstraksi informasi, seperti lokasi. NER dapat diselesaikan dengan menggunakan pendekatan pembelajaran mesin (ML). NER dengan performa yang baik dapat menghasilkan ekstraksi lokasi yang lebih relevan dibandingkan dengan mencari kata pada dataset *gazetterr*. Penelitian terkait NER Bahasa Inggris banyak yang sudah memiliki performa model pembelajaran mesin yang baik, tetapi NER Bahasa Indonesia masih banyak yang perlu ditingkatkan. Hal ini dikarenakan sedikitnya dataset NER berlabel Bahasa Indonesia, yang berdampak terhadap kurangnya metode yang diusulkan.

BERT salah satu algoritma pembelajaran mesin yang dilatih pada tugas – tugas *Natural Language Processing* (NLP) *Next Sentence Prediction* (NSP) dan *Masked Language Modelling (MLM)*. Namun, untuk tugas NER perlu dilakukan *tuning* terlebih dahulu untuk mendapatkan performa lebih baik. Proses *fine-tuning* dapat berupa pelatihan dengan dataset berlabel dan penambahan lapisan saat melakukan proses latih dengan dataset berlabel.

Pada penelitian ini mengusulkan *fine-tuning* BERT dengan mengkombinasikan CNN, Bi-LSTM, Bi-GRU, dan CRF. BERT yang digunakan merupakan beberapa *Pre-Training* model Bahasa Indonesia dan Bahasa Inggris. Dataset pada penelitian ini yaitu NERGrit-IndoNLU, NERGrit-Corpus, NERUGM, dan NERUI. Dari sekenario yang telah ditentukan, total ada 16 model dengan 15 kombinasi ML dan 1 tuning *fully connected layer*. Hasil pengujian memperlihatkan model – model yang diusulkan pada penelitian ini beberapa menghasilkan hasil terbaik pada tiap – tiap dataset. *F1 micro* dataset NERUI 95%, *F1 micro* NERUGM 84%, dan *F1 micro* NERGRIT-Corpus 84%, dan *F1 macro* dataset NERGRIT-IndoNLU 81%. Namun, untuk NER rerata pada model BERT_CNN menghasilkan akurasi terbaik 95.43%, F1 *micro* 80.9%, dan F1 *macro* 76.16%. Performa NER pada model yang digunakan sudah cukup baik sehingga dapat digunakan ekstraksi lokasi yang lebih relevan jika dibandingkan dengan mencari lokasi langsung dengan *gazetteer*.

Kata kunci: Ekstraksi Lokasi, NER, BERT, Pre-Training, Pembelajaran Mesin.