ABSTRACT

The main task of NER is to identify and classify entities in a text into classes that have been determined. Ambiguity is one of the main challenges in identifying named entities in a text, which is to recognize words that have many meanings in different sentences. Some of the approaches that can be made to detect entities are using rule-base, machine learning, and deep learning. At present the approach of using deep learning has gained a lot of success in the NLP field including NER.

In this study context-based deep learning models were built to detect Indonesian-language entities, especially for the conversation dataset and complaints of public services for national health insurance using BERT and comparing the proposed BERT model with the previous model. It also evaluates the effect of using capital letters on NER performance. The pre-processing stage is done by adding BERT masking, token embedding, and positional embedding which are then processed using Transformers with attention approach. Then the model is evaluated by measuring the results of the confusion matrix.

The results showed that BERT multilingual uncased got f1 score 83.4, and BERT multilingual cased with a f1 score 85.4. This shows BERT can achieve better results than the previous method using Bidirectional LSTM-CRF with f1 score 75.6. The results also show that maintaining uppercase letters in the dataset influences the results in NER with better results.

Keywords – Natural Language Processing, Named Entity Recognition, BERT

INTISARI

Tugas utama NER adalah mengidentifikasi dan mengklasifikasikan entitas dalam suatu teks ke dalam kelas yang telah ditentukan. Ambiguitas merupakan salah satu tantangan utama dalam mengidentifikasi entitas bernama dalam sebuah teks, yaitu untuk mengenali kata-kata yang memiliki banyak makna pada kalimat yang berbeda. Beberapa pendakatan yang dapat dilakukan untuk mendeteksi entitas yaitu menggunakan *rule-base*, *machine learning,* dan *deep learning*. Saat ini pendekatan menggunakan *deep learning* mendapatkan banyak keberhasilan di bidang NLP termasuk NER.

Pada penelitian ini dibangun model *deep learning* berbasis konteks untuk mendeteksi entitas berbahasa Indonesia khususnya untuk *dataset* percakapan dan pengaduan layanan publik asuransi kesehatan nasional menggunakan BERT dan melakukan perbandingan model BERT yang diusulkan dengan model sebelumnya. Selain itu juga mengevaluasi pengaruh penggunaan huruf kapital pada performa NER. Tahap pra pemrosesan yang dilakukan yaitu dengan menambahkan *BERT masking, token embedding*, dan *positional embedding* yang kemudian diproses menggunakan *Transformers* dengan pendekatan *attention*. Kemudian model dievaluasi dengan melakukan pengukuran hasil *confusion matrix.*

Hasil penelitian menunjukkan BERT *multilingual uncased* mendapatkan skor f1 83.4, dan BERT *multilingual cased* dengan skor f1 85.4. Hal ini menunjukkan BERT dapat mencapai hasil yang lebih baik dibandingkan metode sebelumnya yang menggunakan *Bidirectional* LSTM-CRF dengan hasil skor f1 75.6. Hasil juga menunjukkan bahwa mempertahankan huruf kapital pada *dataset* mempengaruhi hasil dalam NER dengan hasil yang lebih baik.

Kata kunci – *Natural Language Processing*, *Named Entity Recognition*, BERT