ABSTRACT

PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORK (CNN) AND LONG-SHORT TERM MEMORY (LSTM) METHODS ON DATA CLASSIFICATION OF COVID-19 NEWS TITLE

By

LUTHFIA NUR AZIZAH

The performance of deep learning methods on the classification of text data with different imbalance ratios is an important discussion because the existing data is inherently imbalanced. This study is looking for a reliable deep learning method to classify data on Indonesian news headlines about COVID-19 with several data imbalance ratios. Several data imbalance ratios were made by taking samples from news events using simple random sampling of 30%, 20%, 10%, and 1%. The performance of the CNN and LSTM methods was tested using 10-fold cross validation and compared based on accuracy, precision, recall, and f1-score. The CNN model architecture built in this study generally consists of an input layer, word embedding layer, two convolutional layers, one pooling layer, flatten, two hidden layers, an output layer. A batch normalization layer and a dropout layer after each layer. The LSTM model architecture built in this study generally consists of an input layer, a word embedding layer, two LSTM layers, two hidden layers, an output layer and a dropout layer after each layer. The performance of CNN and LSTM with the Bag of Words (BoW) model as word embedding in this study is quite competitive because CNN outperforms LSTM on all evaluation measures at 37%, 20%, and 10% data imbalance levels, while LSTM outperforms CNN on all evaluation measures at 30% data imbalance levels. Although CNN and LSTM have competitive performance results, LSTM consumes significantly longer computational time than CNN.

Keywords: Classification, Imbalance Data, Deep Learning, CNN, LSTM, K-Fold Cross Validation, Bag of Words (BoW), COVID-19 news headlines.

ABSTRAK

KINERJA METODE *CONVOLUTIONAL NEURAL NETWORK* (CNN) DAN *LONG-SHORT TERM MEMORY* (LSTM) PADA KLASIFIKASI DATA JUDUL BERITA COVID-19

Oleh

LUTHFIA NUR AZIZAH

Kinerja metode *deep learning* pada klasifikasi data teks dengan rasio ketidakseimbangan yang berbeda merupakan diskusi yang penting karena data yang ada pada dasarnya tidak seimbang. Penelitian ini mencari metode *deep learning* yang dapat diandalkan untuk mengklasifikasi data judul berita berbahasa Indonesia tentang COVID-19 dengan beberapa rasio ketidakseimbangan data. Beberapa rasio ketidakseimbangan data dibuat dengan mengambil sampel dari berita *event* menggunakan *simple random sampling* sebanyak 30%, 20%, 10%, dan 1%. Kinerja metode CNN dan LSTM diuji menggunakan 10-*fold cross validation* dan dibandingkan berdasarkan akurasi, presisi, *recall*, dan *f1-score*. Arsitektur model CNN yang dibangun pada penelitian ini secara umum terdiri dari *input layer*, *word embedding layer*, dua *convolutional layer*, satu *pooling layer*, *flatten*, dua *hidden layer*, *output layer* serta *batch normalization layer* dan *dropout layer* berada setelahnya pada setiap *layer* tersebut. Arsitektur model LSTM yang dibangun pada penelitian ini secara umum terdiri dari *input layer*, *word embedding layer*, dua *LSTM layer*, dua *hidden layer*, *output layer* serta *dropout layer* berada setelahnya pada setiap *layer* tersebut. Kinerja CNN dan LSTM dengan model *Bag of Words* (BoW) sebagai *word embedding* pada penelitian ini cukup bersaing karena CNN mengungguli LSTM pada semua ukuran evaluasi pada tingkat ketidakseimbangan data 37%, 20%, dan 10%, sedangkan LSTM mengungguli CNN pada semua ukuran evaluasi pada tingkat ketidakseimbangan data 30%. Meskipun CNN dan LSTM memiliki hasil kinerja yang saling bersaing, namun LSTM menghabiskan waktu komputasi yang jauh lebih lama daripada CNN.

Kata kunci: Klasifikasi, Ketidakseimbangan Data, *Deep Learning*, CNN, LSTM, *K-Fold Cross Validation*, *Bag of Words* (BoW), Judul Berita COVID-19.