Rancang Bangun Model Normalisasi Alamat di Indonesia Berbasis CNN dan BERT = Design of Address Normalization Model in Indonesia Based on CNN and BERT

Gilang Setyawan Yoga Pratama, author

Abstrak

Alamat adalah informasi yang digunakan untuk menunjukkan lokasi suatu tempat. Didalamnya terdapat beberapa komponen seperti nama jalan, nomor rumah nomor rumah, RT/RW, kelurahan, kecamatan, kota/kabupaten, provinsi, dan kode pos. Fungsi alamat sebagai identitas goegrafis suatu tempat yang digunakan sebagai komunikasi, pengiriman barang, administasi dan kepentingan layanan lainnya. Normalisasi alamat merupakan proses yang dilakukan untuk mencapai suatu keseragaman dan akurasi komponen yang ada didalamnya. Dataset akan dibuat sendiri menggunakn teknik web scraping yang akan mengumpulkan alamat dengan bantuan Google Maps. Lalu alamat akan dilakukan praproses sebelum digunakan untuk pelatihan model. Dataset akan dibagi menjadi data train dan data test yang akan digunakan untuk pelatihan dan pengujian model. Penelitian ini berfokus pada pengembangan model machine learning dengan teknik convolutional neural network (CNN) dan bidirectional encoder representation from transformer (BERT). Hasil nantinya akan evaluasi berdasarkan accuracy, precision, recall, dan F1-score. Setelah mendapat model terbaik akan dilanjutkan dengan pengujian pada data test dan pengujian manual melalui terminal. Pengguna dapat mengisi alamat langsung lalu akan diberikan output alamat yang teah di lakukan standrisasi. Solusi yang dikembangkan terbagi menjadi 3 model yaitu model CNN, BERT, dan kombinasi CNN + BERT. Berdasarkan hasil peneletian, Model CNN mendapat hasil akurasi sebesar 89%, BERT mendapat hasi akurasi sebesar 23%, dan kombinasi CNN +BERT mendapat hasil akurasi sebesar 27%. Dengan ini model terbaik yaitu CNN akan dipilih untuk masuk ke pengujian menggunakan data test dan pengujian secara manual di terminal.

.....Address is information used to indicate the location of a place. It contains several components such as street name, house number, RT/RW, sub-district, district, city/regency, province, and postal code. The function of the address as a geographic identity of a place used for communication, delivery of goods, administration and other service interests. Address normalization is a process carried out to achieve uniformity and accuracy of the components contained therein. The dataset will be created independently using web scraping techniques that will collect addresses with the help of Google Maps. Then the address will be preprocessed before being used for model training. The dataset will be divided into train data and test data which will be used for training and testing the model. This research focuses on the development of a machine learning model with convolutional neural network (CNN) and bidirectional encoder representation from transformer (BERT) techniques. The results will later be evaluated based on accuracy, precision, recall, and F1-score. After getting the best model, it will be continued with testing on test data and manual testing via the terminal. Users can fill in the address directly and then will be given the address output that has been standardized. The developed solutions are divided into 3 models, namely the CNN model, BERT, and a combination of CNN + BERT. Based on the research results, the CNN model got an accuracy of 89%, BERT got an accuracy of 23%, and the combination of CNN + BERT got an accuracy of 27%. With this, the best model, namely CNN, will be selected to enter the test using test data and manual testing at the

terminal.