INTISARI

Deteksi Kesamaan Pertanyaan pada Forum *Online* menggunakan *Convolutional Neural Network*

Oleh

Damar Adi Prabowo

13/347539/PA/15293

Forum *online* adalah platform untuk berkumpul, berbagi informasi dan berdiskusi antar pengguna mengenai suatu topik tertentu. Pengguna pada forum *online* dapat bertanya mengenai suatu topik, kemudian pengguna lain yang ahli mengenai topik yang ditanyakan menjawab pertanyaan tersebut. Akan tetapi, karena pengguna dapat menanyakan pertanyaan dengan cara yang berbeda – beda, pengguna terkadang menanyakan pertanyaan yang sebelumnya sudah pernah ditanyakan oleh pengguna lain. Oleh karena itu, diperlukan model untuk mendeteksi kesamaan pertanyaan pada forum *online.*

Pada penelitian ini metode yang digunakan untuk mendeteksi kesamaan pertanyaan yaitu *Convolutional Neural Network* (CNN). Kalimat pertanyaan dijadikan vektor dengan menggunakan *word embedding. Pretrained word embedding* yang digunakan yaitu *GloVe word vectors.* Pasangan pertanyaan yang sudah menjadi *word embeddings* digunakan sebagai masukan CNN yang kemudian dibandingkan kesamaannya dengan *Siamese Neural Networks.* Optimasi model dilakukan dengan menggunakan *Stochastic Gradient Descent.*

Model deteksi kesamaan pertanyaan yang dibuat menghasilkan akurasi sebesar **79%**. Akurasi model CNN pada dataset yang digunakan terbukti lebih tinggi dari model yang dibuat dengan algoritma *Jaccard Similarity* dan *Multilayer Perceptron.* Penggunaan dimensi *word embedding* yang tinggi membutuhkan jumlah *epoch* yang lebih tinggi pada proses pelatihan.

Kata kunci : *Convolutional Neural Network, Siamese Neural Network, Quora online forum, Word Embeddings, GloVe*

ABSTRACT

Online Forum Duplicate Question Detection using Convolutional Neural Network

By

Damar Adi Prabowo

13/347539/PA/15293

Online forums are platforms for gathering, sharing information, and discussing between users on a particular topic. Users in online forums can ask questions about a topic, then other users who are experts on the topic of that question would answer the question. However, because users can ask questions in different ways, users sometimes ask questions that other users have previously asked. Therefore, a model is needed to detect the semantic similarity of questions in online forums.

In this study, the method used to detect the semantic similarity of questions is Convolutional Neural Networks (CNN). The question sentence transformed into a vector using word embeddings. Pretrained word embeddings used is GloVe word vectors. The pair of questions that have transformed into word embeddings are used as input for the CNN then the similarities of the questions pair compared with Siamese Neural Networks. Model Optimization is done using Stochastic Gradient Descent.

The semantic similarity detection model of questions resulted in an accuracy of **79%**. The accuracy of the CNN model on the dataset proved to be higher than models made with the Jaccard Similarity and Multilayer Perceptron algorithms. The use of a higher word embedding dimensions requires a higher number of epochs in the training process.

Keyword : Convolutional Neural Network, Siamese Neural Network, Quora online forum, Word Embeddings, GloVe