INTISARI

Normalisasi Teks Media Sosial Twitter dalam Bahasa Indonesia Menggunakan *Word Embedding*

Oleh

Hardiana
16/398509/PA/17470

Twitter memiliki informasi dalam jumlah yang besar. Namun, informasi ini juga mengandung banyak *noisy* dan kata tidak standar yang menyulitkan untuk memilah informasi yang penting. Karena hal ini, perlu dilakukan normalisasi terhadap informasi yang diperoleh dari Twitter untuk mengubah kata menjadi bentuk standar.

Penelitian sebelumnya menunjukkan hasil yang baik dengan menggunakan word embedding dalam melakukan normalisasi teks. Word embedding dapat menangkap hubungan semantik dari kata. Fitur ini menjadi kelebihan dari word embedding yang akan dimanfaatkan untuk menangkap kesamaan semantik kata dari kata-kata pada Twitter. Pada penelitian ini, terdapat dua macam model word embedding yang dilatih . Model yang dilatih yaitu Word2Vec dan FastText. Penelitian ini ditujukan untuk mengetahui performa masing-masing dalam melakukan normalisasi dan membandingkan keduanya untuk untuk mengetahui yang lebih unggul.

Berdasarkan hasil evaluasi, diketahui performa Word2Vec, performa FastText, dan model terbaik di antara keduanya. Model akhir Word2Vec yaitu model dengan learning rate 0.025, ukuran window 5, epoch 200, dan ukuran vektor 150. Hasil evaluasi menunjukkan model ini memperoleh akurasi sebesar 42.06% dan f1-score sebesar 41.87%. Model akhir FastText yaitu model dengan learning rate 0.025, ukuran window 7, jumlah epoch 300, dan ukuran vektor 300. Akurasi yang diperoleh yaitu 59.08% dan f1-score sebesar 68.14 % . Dari perbandingan model akhir keduanya diperoleh model FastText unggul dibandingkan model Word2Vec dalam melakukan normalisasi dengan menggunakan data di luar corpus. Untuk data pada corpus dengan pengambilan secara random Model FastText juga unggul dibandingkan model Word2Vec. Word2Vec memperoleh akurasi 59.41% dan f1-score 66.86% sedangkan FastText memperoleh akurasi 66.02% dan 75.68% pada f1-score. Untuk data pada corpus dengan frekuensi tertinggi, Model Word2Vec unggul dibandingkan model FastText. Word2Vec memperoleh akurasi 77.39% dan f1-score 84.38% sedangkan FastText memperoleh akurasi 69.89% dan f1-score 78.30%.

Kata-kata kunci : Twitter, normalisasi teks, word embedding, Word2Vec, FastText.

ABSTRACT

Text Normalization on Indonesian Twitter Data Using Word Embedding

By

Hardiana
16/398509/PA/17470

Twitter has a large amount of information. This information contains noise and non-standard words that make it difficult to collect the essential information in it. Therefore, it is necessary to normalize the information obtained from Twitter in order to convert words into standard forms.

Previous studies have shown good results by using word embedding in normalizing text. Word embedding is a method of word representation that converts words into vector numbers. Word embedding can capture the semantic relationships of words. This feature is one of the advantages of word embedding that will be used to capture the semantic similarity from words on Twitter. In this study there are two types of word embedding models that are trained using Twitter data. The trained models are Word2Vec and FastText. This study aims to compare both models to see the performance of each in normalizing and to find out which model performs better.

Based on the evaluation results, we present the Word2Vec's performance, FastText's performance, and the best model of these two models. The final model of Word2Vec is a model with a learning rate of 0.025, window size 5, number of epoch is 200, and the vector size is 150. Evaluation results show that this model report 42.06% accuracy and 41.87% f1-score. The final model for FastText is a model with 0.025 learning rate , window size of 7, number of epoch is 300, and the vector size of 300. This model reached 59.08% accuracy and 68.14% f1-score. From the comparison between the best models, the FastText model is superior to the Word2Vec model in normalizing using data outside the corpus. For data on corpus with random retrieval, the FastText model is also superior to the Word2Vec model. Word2Vec obtained an accuracy of 59.41% and f1-score of 66.86% while FastText obtained an accuracy of 66.02 % and 75.68 % on the f1-score. For data on the corpus with the highest occurrence, the Word2Vec Model is superior to the FastText model. Word2Vec obtained an accuracy of 77.39 % and an f1-score of 84.38 %, while FastText obtained an accuracy of 69.89 % and an f1-score of 78.30 %.

Keywords : Twitter, text normalization, word embedding, Word2Vec, FastText.