

INTISARI

Perbandingan Kinerja Beberapa Metode *Word Embedding* pada Bahasa Indonesia

Oleh

Muhammad Farhan Fauzan Mulyana
18/430268/PA/18781

Word embedding adalah representasi vektor bernilai riil dari kata-kata dengan menyematkan makna semantik dan sintaksis yang diperoleh dari korpus besar yang tidak berlabel. *Word embedding* sangat umum digunakan untuk menyelesaikan permasalahan *natural language processing*. *Word embedding* terus mengalami perkembangan dan juga bermunculan algoritma terbaru dengan model arsitektur yang berbeda dari sebelumnya. Banyaknya *word embedding* yang tersedia memberi banyak pilihan saat membangun suatu arsitektur. Berkaitan dengan hal tersebut diperlukan suatu perbandingan untuk menentukan model *word embedding* yang paling tepat digunakan sebagai bagian dari arsitektur *natural language processing*.

Penelitian ini membandingkan kinerja dan lama waktu *training* dari *word embedding* Word2Vec, GloVe, fastText, dan BERT. Perbandingan dilakukan dengan menggunakan dua topik yang diselesaikan yaitu analisis sentimen dan *named entity recognition*.

Hasil penelitian ini menunjukkan bahwa model BERT menghasilkan solusi terbaik namun dengan waktu *training* terlama. Model Word2Vec, GloVe, dan fastText memiliki kemiripan pada waktu *training* dengan selisih rata-rata 70 milidetik. Model Word2Vec menghasilkan kinerja terbaik kedua dan tercepat ketiga. Model GloVe menghasilkan kinerja terbaik ketiga dan tercepat kedua. Model fastText menghasilkan kinerja terburuk namun menjadi yang paling cepat di antara semua model.

Kata kunci: perbandingan algoritma, *word embedding*, Word2Vec, GloVe, fastText, BERT, analisis sentimen, *named entity recognition*

ABSTRACT

Performance Comparison Of Some Word Embedding Methods in Indonesian Language

Oleh

Muhammad Farhan Fauzan Mulyana
18/430268/PA/18781

Word embedding is a real-valued vector representation of words by embedding the semantic and syntactic meanings derived from a large unlabeled corpus. word embedding continues to develop and new algorithms emerge with different architectural models than before. The large number of word embedding available gives us many choices when building an architecture. In this regard, a comparison is needed to see which word embedding model is most appropriate to use as part of the natural language processing architecture.

This study compares the performance and training time from Word2Vec, GloVe, fastText, and BERT. Comparisons were made using two such as sentiment analysis and named entity recognition.

The result of this research shows that BERT model produces the best solution but with the longest training time. Word2Vec, GloVe, and fastText models are similar in training time with an average difference of 70 milliseconds. The Word2Vec has the second best performance with the third fastest time. The GloVe model has the third best performance with the second fastest time. The fastText model has the worst performance with the fastest time in training.

Kata kunci: perbandingan algoritma, *word embedding*, Word2Vec, GloVe, fastText, BERT, analisis sentimen, *named entity recognition*