INTISARI

PERBANDINGAN WORD EMBEDDING UNTUK ANALISIS SENTIMEN PADA DATA TWITTER

oleh

Ainun Nafilatur Rusyda

15/379615/PA/16673

Twitter merupakan platform yang menyediakan banyak data teks singkat mengenai opini mengenai berbagai macam hal. Data twit banyak digunakan untuk diolah menjadi informasi sentimen terhadap suatu hal dengan sentimen analisis. Jenis data ini memiliki kekurangan yaitu *noise* dan memiliki fitur semantik unik karena ukurannya yang pendek. Oleh karena itu pemilihan fitur dan model analisis sentimen menjadi hal yang penting untuk menentukan kualitas hasil sentimen pada data teks singkat.

Penelitian ini mengimplementasikan metode word embedding untuk pembuatan fitur analisis sentimen. Word embedding merupakan metode representasi kata berbentuk vektor bernilai riil dengan dimensi rendah yang disebut vektor embedding. Vektor embedding dibuat dengan pembelajaran dari data teks yang besar untuk dapat menangkap sintaks dan semantik kata. Terdapat beberapa penelitian yang menyediakan vektor embedding yang sudah melalui proses pembelajaran (*pre-trained)* yaitu Word2Vec dan Glove. Kedua vektor embedding akan digunakan dalam penelitian ini bersama dengan vektor embedding yang dibuat oleh Quanzi. Vektor embedding tersebut akan digunakan untuk pembuatan fitur. Setelah itu, fitur yang dihasilkan akan dibandingkan menggunakan algoritma SVM, LSTM dan GRU.

Pengujian dilakukan dengan menghitung nilai akurasi, *precision*, *recall* dan skor *F-1*. Berdasarkan penelitian yang telah dilakukan, ketiga model klasifikasi memberikan hasil akurasi tertinggi sebesar 79% terhadap dua dataset yang digunakan. Untuk vektor embedding buatan Quanzi dan Word2Vec memberikan peningkatan 1% akurasi dari vektor embedding Glove.

Kata kunci: Sentimen Analisis, Word Embedding, Word2Vec, Glove, SVM, LSTM, GRU

ABSTRACT

COMPARISON OF WORD EMBEDDINGS FOR SENTIMENT ANALYSIS OF TWITTER DATA

by

Ainun Nafilatur Rusyda

15/379615/PA/16673

Twitter is a platform providing a lot of short text data of opinions on many different things. Tweets data are often processed to be underlying sentiment information to a certain issue by using a sentiment analysis. However, this type of data also has its shortcomings as it contains noise and its semantic features are unique due to its short size. Therefore, selection on features and sentiment analysis model become important as to determine the quality of sentiment results in short data texts.

This research employs word embedding method to obtain features of sentiment analysis. Word embedding is a method of word representation in a form of real-valued vector with low dimension, which also known as embedding vector. This embedding vector is obtained through learning from large text corpus as to capture syntax and semantics of the word. Some researches actually provide pre-trained vector embeddings such as Word2Vec and Glove. These two embeddings will be used in this research along with vector embedding produced by Quanzi. These vector embeddings will be employed to obtain features. In what follows, the resulted features are compared to SVM, LSTM, and GRU algorithms.

The testing was done by calculating the accuracy value, precision, recall and F-1 score. Based on the conducted research, the highest accuracy resulted by three models of classifications to the two used datasets is on 79%. For vector embedding managed by Quanzi and Word2Vec, they give 1% accuracy increasing from Glove embedding vector.

Keyword: Sentiment Analysis, Word Embedding, Word2Vec, Glove, SVM, LSTM, GRU