

INTISARI

SENTIMENT ANALYSIS PADA TWEET BERBAHASA INDONESIA MENGUNAKAN WORD2VEC EMBEDDING DENGAN METODE KLASIFIKASI SVM DAN NAÏVE BAYES

Emerald Taufik Akbar Makerti

16/398501/PA/17462

Dengan jumlah pengguna aktif sebesar 14.8 juta, netizen Indonesia adalah satu pengguna terbesar media sosial Twitter. Jutaan *tweet* yang berisikan pikiran dan opini dikirim tiap harinya dalam media sosial Twitter. Maka dari itu, diperlukan otomasi analisa sentimen berbasis *Natural Language Processing*. Namun, penggunaan ekstraksi fitur *one hot vector* tidak bisa menyimpan makna semantik dari kata. Maka dari itu digunakan *word embedding* agar model dapat mengenali makna semantik dari kata.

Penelitian ini berfokus untuk mengembangkan suatu model analisa sentimen berbasis Word2Vec *embedding* dengan metode klasifikasi SVM dan Naïve Bayes. Pada penelitian ini dataset yang digunakan adalah kumpulan *tweet* dalam bahasa Indonesia.

Pada penelitian ini kombinasi Word2Vec embedding dengan metode klasifikasi SVM dan Naïve Bayes mampu menghasilkan tingkat akurasi yang cukup tinggi. Arsitektur kombinasi Word2Vec berdimensi 300 dan SVM menghasilkan tingkat akurasi prediksi sentimen sebesar 76,7%, *precision* sebesar 70,9%, *recall* sebesar 63,3%, dan *F1-score* sebesar 66,9%. Model Naïve bayes menghasilkan akurasi sebesar 72,2%, *precision* sebesar 64%, *recall* sebesar 57,2%, dan *F1-score* sebesar 60,4% dengan ekstraksi fitur Word2Vec berukuran 100 dimensi.

Kata kunci: Analisa sentimen, Word Embedding, Word2Vec, Support Vector Machine, Naïve Bayes, Natural Language Processing

ABSTRACT

SENTIMENT ANALYSIS OF INDONESIAN TWEETS USING WORD2VEC EMBEDDING WITH SVM AND NAÏVE BAYES CLASSIFICATION METHOD

Emerald Taufik Akbar Makerti
16/398501/PA/17462

With active users clocking in at 14.8 million users, Indonesian netizens are one of the most active users of Twitter social media. Millions of *tweets* are sent daily to the Twitter platform containing thoughts and opinions. Therefore, there is a need to automate sentiment analysis using Natural Language Processing (NLP). However, feature extraction method using one hot vector will render words semantic meaning to be lost. That's why we need to use word embedding to keep word semantic meaning recognizable by the model.

This study will focus on developing a sentiment analysis model using Word2Vec embedding with SVM and Naïve Bayes classification method. The dataset that will be used consists of *tweets* in Indonesian language.

In this study, the combination of Word2Vec embedding with 300-dimension size and SVM produced performance metrics of 76.7% accuracy, 70.9% precision, 63.3% recall, and 66.9% F1-score. While combination of Naïve Bayes and 100-dimension size Word2Vec produced performance metrics of 72.7% accuracy, 64% precision, 57.2% recall, 60.4% F1-score.

Keywords: Sentiment analysis, Word Embedding, Word2Vec, Support Vector Machine, Naïve Bayes, Natural Language Processing