Klasifikasi Sentimen pada Dataset Ulasan Film menggunakan *Machine Learning* dan OpenAI *Text Embedding*

Azzam Abdurrahman[1], Moch. Arif Bijaksana[2], Kemas Muslim Lhaksmana[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]azzamabdurrahman@student.telkomuniversity.ac.id,
[2]arifbijaksana@telkomuniversity.ac.id, [3]kemasmuslim@telkomuniversity.ac.id

Abstrak

Analisis sentimen pada ulasan film menjadi semakin penting seiring dengan meningkatnya volume data tekstual. Performa model *machine learning* untuk tugas ini sangat bergantung pada kualitas representasi teks yang digunakan. Penelitian ini bertujuan untuk mengevaluasi efektivitas model *embedding* teks kontekstual dari OpenAI, Text-embedding-3-large, untuk klasifikasi sentimen pada dataset *Movie Reviews*. Metodologi penelitian mencakup dua pendekatan klasifikasi: *supervised learning* menggunakan *Support Vector Machine* dan *Logistic Regression*, serta klasifikasi *zero-shot*. Performa Text-embedding-3-large dibandingkan secara langsung dengan model *embedding* statis Word2Vec pada dataset yang telah dibersihkan dan dataset asli. Hasil penelitian menunjukkan bahwa Text-embedding-3-large secara signifikan mengungguli Word2Vec, dengan peningkatan *F1-score* dari 78.01% menjadi 93.20%. Konfigurasi terbaik dicapai oleh kombinasi *Support Vector Machine* dengan *hyperparameter default* pada dataset yang tidak dibersihkan, yang mengindikasikan kemampuan model memanfaatkan informasi kontekstual dari tanda baca. Selain itu, pendekatan *zero-shot* menunjukkan kinerja yang cukup baik dengan *F1-score* 86.29%, yang membuktikan kapabilitas generalisasi model tanpa memerlukan data latih berlabel.

Kata kunci : klasifikasi sentimen, ulasan film, *machine learning*, openai, *embedding* teks, zero-shot.

Abstract

Sentiment analysis on film reviews is becoming increasingly important as the volume of textual data grows. The performance of machine learning models for this task is highly dependent on the quality of the text representation used. This research aims to evaluate the effectiveness of OpenAI's contextual text embedding model, Text-embedding-3-large, for sentiment classification on the Movie Reviews dataset. The research methodology includes two classification approaches: supervised learning using Support Vector Machine and Logistic Regression, as well as zero-shot classification. The performance of Text-embedding-3-large is directly compared with the Word2Vec static embedding model on both cleaned and uncleaned dataset. The results show that Text-embedding-3-large significantly outperforms Word2Vec, with an F1-score increase from 78.01% to 93.20%. The best configuration is achieved by the combination of Support Vector Machine with default hyperparameter on the uncleaned dataset, which indicates the model's ability to utilize contextual information from punctuation. Furthermore, the zero-shot approach shows quite good performance with an F1-score of 86.29%, which proves the model's generalization capabilities without requiring labeled training data.

Keywords: sentiment classification, movie reviews, machine learning, openai, text embedding, zero-shot