INTISARI

PENGEMBANGAN MODEL BERT DAN BACK TRANSLATION UNTUK ANALISA SENTIMEN CODEMIXED PADA DATA TWITTER

Oleh

Nisrina Hanifa Setiono

23/525832/PPA/06622

Rendahnya performa model BERT dalam analisis sentimen pada data code-mixed Twitter yang bersifat informal disebabkan oleh penggunaan slang, singkatan, kata tidak baku, serta inkonsistensi ejaan dan tata bahasa yang tidak sesuai dengan karakteristik data pretraining formal, sehingga menyebabkan akurasi analisis sentimen masih rendah. Tantangan tersebut menjadi permasalahan penting dalam bidang pemrosesan bahasa alami (NLP), khususnya untuk kombinasi Bahasa Indonesia-Inggris yang banyak digunakan di media sosial.

Penelitian ini bertujuan untuk meningkatkan performa model BERT dalam menganalisis sentimen pada dataset code-mixed dengan mengembangkan kombinasi antara model BERT dan teknik back translation. Pendekatan ini dirancang khusus untuk mengatasi tantangan linguistik pada data informal code-mixed Bahasa Indonesia-Inggris, sehingga diharapkan mampu meningkatkan akurasi dalam analisis sentimen. Metode yang diusulkan diterapkan pada dataset INDONGLISH yang terdiri dari 5.067 cuitan Twitter berlabel sentimen positif, negatif, atau netral.

Hasil penelitian menunjukkan bahwa penerapan back translation secara langsung pada data tweet memberikan hasil lebih baik karena mampu mempertahankan makna asli kalimat, sehingga meningkatkan performa model BERTweet dari 0.7270 menjadi 0.7457. Sebaliknya, ketika back translation diterapkan setelah translasi monolingual, akurasi model justru menurun pada BERTweet dari 0.7270 menjadi 0.7161. Proses translasi berulang menyebabkan perubahan struktur dan konteks kalimat yang signifikan, sehingga label sentimen menjadi kurang sesuai. Temuan ini memperkuat bahwa setiap tambahan proses translasi memiliki risiko menurunkan akurasi analisis sentimen, terutama pada dataset code-mixed yang sangat sensitif terhadap perubahan linguistik.

Kata Kunci: Analisis Sentimen, *Code-Mixed*, *BERT*, *Pretrained Model*, *Back translation*, *Text Style Transfer*

ABSTRACT

DEVELOPMENT OF BERT AND BACK TRANSLATION MODELS FOR SENTIMENT ANALYSIS OF CODEMIXED DATA ON TWITTER

Nisrina Hanifa Setiono

23/525832/PPA/06622

The low performance of the BERT model in sentiment analysis on informal code-mixed Twitter data is due to the use of slang, abbreviations, non-standard words, and inconsistencies in spelling and grammar, which do not match the characteristics of formal pretraining data. This mismatch makes it difficult for the model to accurately understand sentence contexts, resulting in low sentiment analysis accuracy. Addressing this challenge is important in the field of natural language processing (NLP), especially for the combination of Indonesian-English commonly used on social media.

This research aims to improve the performance of the BERT model in analyzing sentiment on code-mixed datasets by developing a combination of the BERT model and back translation techniques. This approach is specifically designed to overcome linguistic challenges in informal Indonesian-English code-mixed data, thereby enhancing the accuracy of sentiment analysis. The proposed method was applied to the INDONGLISH dataset consisting of 5,067 Twitter tweets labeled as positive, negative, or neutral sentiments.

The results show that applying back translation directly to tweet data produced better results by preserving the original meaning of the sentences, thereby increasing the performance of the BERTweet model from 0.7270 to 0.7457. Conversely, applying back translation after monolingual translation reduced the accuracy of the BERTweet model from 0.7270 to 0.7161. Repeated translation processes significantly altered sentence structure and context, resulting in mismatched sentiment labels. These findings indicate that additional translation steps can negatively impact the accuracy of sentiment analysis, particularly on code-mixed datasets that are highly sensitive to linguistic variations.

Keywords: Sentiment Analysis, *Code-Mixed*, *BERT*, *Pretrained Model*, *Back translation*, *Text Style Transfer*