INTISARI

CODE-MIXED SENTIMENT ANALYSIS BERBASIS TRANSFORMER UNTUK DATA MEDIA SOSIAL TWITTER

Laksmita Widya Astuti
21/476104/PPA/06149

Media sosial identik dengan penggunaan bahasa tidak baku. Bahasa tidak baku atau yang disebut juga *slang* bertujuan untuk memperkaya bahasa. Media sosial lekat dengan percampuran atau penyerapan dari beberapa bahasa yang disebut juga dengan *code-mixing*. Penelitian *code-mixing* khususnya analisis sentimen masih menjadi tantangan tersendiri di ranah NLP. Kurangnya kumpulan data *code-mixed* Bahasa Indonesia-Inggris beranotasi serta keterbatasan beberapa metode *embedding* yang berfokus hanya pada 1 bahasa dapat menyebabkan performa yang buruk dalam hal analisis sentimen karena tidak mampu memahami konteks serta memiliki kata yang *out-of-vocabulary* atau OOV.

Pretrained model yang dilatih pada korpus besar mampu mempelajari representasi bahasa secara universal. *Pretrained model* bermanfaat untuk menyelesaikan tugas di bidang NLP dan menghindari pelatihan model baru dari awal. *Bidirectional Encoder Representations from Transformers* (BERT) merupakan salah satu metode *pretraining* yang dapat merepresentasikan bahasa. BERT dapat melatih model dengan pemahaman bahasa tertentu pada korpus yang besar seperti Twitter. Penelitian terdahulu menghasilkan *pretrained model* Bahasa Indonesia dan Inggris yang dilatih pada data *monolingual* namun tidak dilatih pada data *multilingual* secara langsung. Dengan keterbatasan tersebut, 5 skenario dilakukan dalam penelitian ini untuk dibandingkan dan dievaluasi agar dapat menyelesaikan tugas sentimen analisis pada data *code-mixed* yang sudah dibangun dengan menggunakan *Feed Forward Neural Network*. Data *code-mixed* Bahasa Indonesia-Inggris yang dibangun diannotasi oleh 5 *annotator* dari beberapa keilmuan seperti sastra Indonesia, sastra Inggris, serta *data science*. Pengambilan data dilakukan dengan menggunakan teknik *majority vote*.

Dari dataset yang sudah dibangun dan dievaluasi melalui kelima skenario tersebut, hasil menunjukkan bahwa *pretrained model* IndoBERTweet mampu mencapai hasil yang terbaik dalam menyelesaikan tugas analisis sentimen melalui penambahan skenario *preprocessing* dan *pretrained model* yang berbeda pada data *code-mixed* Bahasa Indonesia-Inggris dengan nilai performa rata-rata *precision* sebesar 76.07%, *recall* 75.52%, *f-1 score* 75.51%, dan akurasi sebesar 76.56%.

Kata Kunci: Analisis Sentimen, *Data code-mixed*, *BERT, Pretrained Model, Google Translate*

ABSTRACT

CODE-MIXED SENTIMENT ANALYSIS USING TRANSFORMER FOR TWITTER SOCIAL MEDIA DATA

Laksmita Widya Astuti
21/476104/PPA/06149

Social media is synonymous with the use of non-standard language. Non-standard language or what is also called slang aims to enrich the language. Social media is closely related to mixing or absorbing several languages, also known as code-mixing. Code-mixing research, especially sentiment analysis, is still a challenge in the realm of NLP. The lack of annotated Indonesian-English code-mixed dataset and the limitations of some embedding methods that focus on only 1 language can lead to poor performance in terms of sentiment analysis because they are unable to understand the context and have out-of-vocabulary or OOV words.

Pretrained models that are trained on a large corpus are able to learn language representations universally. The pretrained model is useful for completing assignments in the field of NLP and avoids training a new model from scratch. Bidirectional Encoder Representations from Transformers (BERT) is a pretraining method that can represent language. BERT can train models with a specific language understanding on a large corpus such as Twitter. Previous research resulted in pre-trained models of Indonesian and English which were trained on monolingual data but not directly trained on multilingual data. With these limitations, 5 scenarios were carried out in this study to be compared and evaluated in order to complete the sentiment analysis task on code-mixed data that has been built using the Feed Forward Neural Network. The Indonesian-English code-mixed data that was built was annotated by 5 annotators from several disciplines such as Indonesian literature, English literature, and data science. Data collection was carried out using the majority vote technique.

From the dataset that has been built and evaluated through the five scenarios, the results show that the IndoBERTweet pre-trained model is able to achieve the best results in solving sentiment analysis tasks by adding different preprocessing and pretrained model scenarios to Indonesian-English code-mixed data with an average precision of 76.07%, recall of 75.52%, f-1 score of 75.51%, and accuracy of 76.56%.

Keywords: Sentiment Analysis, Data code-mixed, BERT, Pretrained Model, Google Translate