ABSTRACT

Social media users nowadays tend to use code-mixed language to express their opinion. The users of social media has exponentially risen in some countries like Indonesia, it has given rise to large volumes of code-mixed data, in which users use more than one language in a single text. Data with code-mixed is often noisy and most importantly, the monolingual model usually does not work well on it. This has been a challenge for Natural Language Processing (NLP) for processing and analyzing the data. In this work, we conduct experiment of sentiment analysis on English-Indonesian code-mixed data. The approach that is by utilizing a multilingual pre-trained model, mBERT. By analyzing the sentiment analysis models' predictions, we may assess how effectively the model can adjust to the implicit noises inherent in code-mixed data. The classification model's performance was tested using batch size and epochs parameters to discover and obtain the highest accuracy. The experimental result shows that the highest accuracy we obtained from the mBERT model that is trained with our dataset obtained was 76\%, with 16 batch size and epochs used 7.

Index Terms: Code-mixed, Sentiment analysis, mBERT, Natural Language Processing