

***Abstract***—In this work, we conduct sentiment analysis on Indonesian-Sundanese code-mixed tweets. Sundanese is one of Indonesia’s regional languages with over 42.000.000 speakers. We use a pre-trained language model, IndoBERT, to tackle the sentiment analysis task. Our evaluation result shows that the best accuracy is 81%. We analyze the errors and find that most mislabeled tweets are because the words on the wrongly predicted tweet contain many words from other labels. It is also possible that it happens since the sentence in the tweet is ambiguous, the words used in the tweet are unavailable in the training data set, or the use of abbreviated words in the tweet.

***Index Terms***—sentiment analysis, IndoBERT, code-mixed data, Indonesian-Sundanese code mixed tweets, natural language processing