1. INTRODUCTION

In Indonesia, there is an actress named Cinta Laura Kiehl or well known as Cinta Laura. She was well known for her English accent when she speaks bahasa Indonesia, also because she has a mixed race Indonesia-Germany makes it difficult for her to speaks Indonesian fluently, so she always speaks mix between Indonesia-English when speaking. While the style of that code-mixing remains iconic in Indonesia even 10 years after the appearance of Cinta Laura, code-mixing phenomenon is common in multilingual countries like Indonesia. The practice of incorporating linguistic constructions from one language, such as words and phrases, into another is known as code-mixing[1]

Social media such as Facebook and Twitter were famous over the past decade. Social media users have exponentially risen in some countries like Indonesia. This has also led to the big amount of code-mixing usage on each platform. It is a common phenomenon that occurs in multilingual communities, and it can take many different forms. For example, a person might use words or phrases from one language in the midst of speaking or writing in another language, or they might switch back and forth between languages within a single sentence or phrase. Code-mixing can also involve combining elements of different language varieties, such as combining standard and colloquial forms of a language or mixing dialects. Code-mixing is a natural and common part of multilingual communication, and it can serve various purposes, such as emphasizing a point, expressing solidarity with a particular group, or adding emphasis or nuance to a message. This has been a challenge for Natural Language Processing (NLP) for processing and analyzing the data. Code-mixed text is a challenge for the NLP practitioners community [2].

In the field of sentiment analysis and transliteration, code-mixing is now popular. It is challenging to use the right method to extract the correct sentiment from code-mixed data [3]. The social text data that is Hindi-English code-mixed have been studied by Vijay et al. [4]. Code-mixing text inherits the vocabulary and grammar of multiple languages and often forms new structures based on the user preference. This poses a challenge to sentiment analysis, as traditional semantic analysis approaches do not capture the meaning of sentences. The lack of annotated data available for sentiment analysis model, the presence of multiple languages in a single text document, and the need to handle language-specific characteristics, such as idioms and collocations also limits progress in this area. There are several techniques to do sentiment analysis on code-mixed data, such as Word2Vec, FastText, Convolutional Neural Network (CNN), multilingual BERT (mBERT), etc.

In this study, we focus on the code-mixed Indonesian-English text. Code-mixing has been referred to as intrasentential code-mixed [5]. The two or more languages were mixed between sentences to make a whole sentence, and can be merged with affixes to make the sentence more meaningful. There has been research for code-mixed Indonesia-English that was conducted in 2020 using word embedding, and the best result accuracy for the study is 67.27\% for Code-Mixed Embedding [6]. Based on that study, Code-Mixed word embeddings also has good potential also for other NLP tasks that require cross-lingual or multilingual word embeddings, in this paper we explore more about code-mixed sentiment analysis using multilingual BERT (mBERT).