I. Introduction

Social media has been the perfect place for anyone, regarding their personalities or background, to express themselves freely. Some even use more than one language in a sentence to communicate with others. By mixing two or more languages, people can express their ideas much more precisely and convincingly. Hence, that is why it is quite common to see people who know more than one language switch from one language to another [1]. Of course, this also happens on social media, especially among Indonesians who usually master more than one language, that is the Indonesian language and their regional language.

This creates challenges for natural language processing since a sentence might involve two languages or more at the same time. One of the reasons is that a word from a language might not have a specific translation in the other language. It is possible to affect the system's effectiveness [2]. But, this also gives an opportunity to explore more types of data that can be processed. In 2021 research shows that a Spanish and English multilingual model outperformed a model that knows the language in which the text is written and a model that acts by what the language identification tool result [3]. Another problem is that the data set for code-mixing corpora are rare, especially for Indonesian and its regional languages.

In 2015, the second most used regional language in Indonesia is Sundanese, with a number of 42,000,000 people, making it the second most used regional language in Indonesia [4]. The use of Sundanese language in social media has been done by a lot of people, this includes several people that code-mixed it with Indonesian language. The use of code-mixing between main language and local language has been done previously in [5], using Hindi-English and Bengali-English code-mixed data set. In [5], the author shared an information regarding a task of sentiment analysis of Indian language at ICON 2017 [6]. Among the registered participants, the best performing team uses word and character level n-grams as features and SVM for sentiment classification. Knowing that research on Indonesian and regional language code-mixed sentiment analysis is still lacking, it can be concluded that more research and data set on this topic is needed.

An example of Indonesian-Sundanese code-mixed data that are frequently used in social media is ”*Lama-lama aing meninggal kerja gini*.” which means ”I will soon die working like this.” in which the word *aing* is Sundanese of ”I” and the rest is in Indonesian. The sentence mentioned before are commonly found on social media, whether in Facebook, Instagram, Twitter, etc. By conducting this study, the potential to understand more text data in social media, specifically with Indonesian-Sundanese code-mixed data, is broadened.

By conducting this study, we can easily understand more Indonesian-Sundanese code-mixed tweets or data relatively quickly. This breakthrough will ease companies understand their customers or consumer, especially in Indonesia, since as time goes by the use of Sundanese-Indonesian code-mixed data or any Indonesian code-mixed data is booming on social media.

In this study, we use Indonesian pre-trained model, IndoBERT. IndoBERT was trained from Indonesian data set collected from public sources such as blogs, websites, news, and media text. It might contain text written in Indonesia's regional language, including Sundanese. A study regarding language models towards Sundanese data set has been done before [7], in which the author compares existing multilingual models with monolingual models. The author also trained a monolingual model for Sundanese dataset, resulting a slightly higher performance than IndoBERT. The author also suspect that IndoBERT might as well applicable as an alternative to multilingual models for Indonesian regional languages. Therefore, we explore and analyze the use of IndoBERT on Indonesian-Sundanese code-mixed tweets.

The author's contribution to this paper is to collect Indonesian-Sundanese code-mixed tweets that have been crawled and checked manually. A total of 786 tweets are collected. The tweets are then labeled manually.