1. Introduction

Social media usage has increased in recent years, and it has become one of the most important components of daily life in the world. The purpose of social media initially was to make connections between people who were not connected. However, social media has become more effective and has more features to improve the lives of people in recent years, especially during the COVID-19 pandemic, regulations imposed to create lockdown. So-cial media such as Twitter, Facebook, Instagram, WhatsApp, and YouTube became the primary sites for society to interact, share opinions, and other information. As the result, variety of social media data are collected from each account users. Based on Statista [1] Twitter is one of the most popular social media platforms with 322.4 million users. Therefore, the researchers are interested to study on media social user personality.

According to this study [2] Personality is a form of individual human thought patterns based on traits, feel-ings, and behavior that distinguishes between each human characteristic. As a psychological construct, the concept of personality refers to characteristics that can be quantified or explained quantitatively and predict observable differences in behavior.

The information that obtained from social media is based on the large amounts of interacted data which can be analyzed under Big Five Personality Prediction on Twitter, aiming to learn about the personality traits of the media social user. Users can be classified by OCEAN (Openness, Conscientiousness, Extraversion, Agreeable-ness, and Neuroticism) based on the word used by users while doing activities on Twitter such as reaction, cap-tion, opinion, and argument. As the result, information is provided about the behavior of each person on social media, to help build a stronger community among the users and they can interact more effectively with others.

In this research, we build a personality prediction system based on Twitter using the Artificial Neural Net-work (ANN) as a classification method of personality traits and determine the accuracy of the predictions fo-cuses on Bahasa Indonesia as the system prediction language. To overcome the imbalance data that can cause overfit while training the data, we use SMOTE technique. With the Linguistic Inquiry Word Count (LIWC) as linguistic feature word counts, it can increase the performance based on the correlation between language and psychology-relevant of the text. RoBERTa (Robustly Optimized BERT pre-training Approach) is also used as a semantic approach to increase the performance of the system built [3]. And finally, Hyperparameter Tuning is used to find the best parameters for the ANN method to achieve the best performance [4].

The purpose of this research is to find out on how to improve the accuracy of the Big Five Personality classification and predictions with word embeddings using RoBERTa as semantic approach. RoBERTa attempts to robust the model by the dynamic masking pattern and improve the performance of the model with larger batch data.

As for the remainder of the paper, section 2 provides a brief overview of related work. In section 3, we present our methodology of personality prediction by five scenario experiments. In section 4, we cover the details of our experiment results and followed by conclusion in section 5.