Bab I Introduction

In the world of work, working in groups can make it easier to complete a job. Knowing the personalities of other group members will be able to help maximize the work of the group. Personality can be defined by how a person interacts with the surrounding environment [1]. Personality comes from the Greek word "*persona*" which means a symbol that represents a person's identity [2]. Personality traits can be seen through a person's thoughts, feelings, and behavior patterns to respond to certain circumstances (Roberts 2009 p 140) [3]. According to Lewis Goldberg [1], there are five main personalities: Openness, Continuousness, Agreeableness, Extraversion, and Neuroticism, commonly abbreviated as OCEAN. Each of these personalities has its advantages.

Today, communicating is easier because of the online platform called social media. Twitter, Facebook, Instagram, WhatsApp, and many more are social media that people use to communicate online. According to we are social media [4], Twitter became one of the most popular social media used among users aged 16 to 64 years in Indonesia in January 2013. Twitter is one of the social media that provides microblogging services that allow users to send and read messages up to 140 characters in one letter called tweets [5]. There is no limit if someone wants to write tweets so that someone can freely express what they want to share.

Several studies try to examine a person's personality through the classification method of the words of Twitter users' tweets. One of them is research done by Willy et al. [6]., who tried to use the Term Frequency Inverse Relevance Frequency method to convert the word tweets into vectors and then use the Decision Tree C.45 method to classify Big Five Personality. This study produces a model that can predict a person's personality by 65.72%. The author revealed that the data used in this study contained dominant data labels, so the model detected more dominant data labels than the others.

Another research was conducted by Salsabila et al. [7]. Who used a dataset of 295 users and 511,617 tweets using the Synthetic Minority Over Sampling Technique (SMOTE), Linguistic Inquiry Word Count (LIWC), and Bidirectional Encoder from Transformers Representations (BERT) methods which resulted in an accuracy of 80.07%. The author reveals that the semantic approach can produce better accuracy because the previously trained BERT model is more applicable to understanding words in sentences. The weakness of the research mentioned by the author is that the dataset is still tiny, namely 295 Twitter users.

Research conducted by Gita et al. [8]. Used the SVM model combined with the TF-IDF, LIWC, and Hyperparameter tuning model to detect Big Five Personality. TF-IDF and LIWC function as feature extraction. Hyperparameter tuning is used to help the model combine various possible parameters to obtain the best parameters, which will later be applied to the model. The results of this study resulted in a baseline accuracy of 74.44%, and when the Hyperparameter Tuning is used to baseline, it gains accuracy to 84.22%. The weakness of this research is the small dataset used by the author, which is 287 Twitter user data.

Research conducted by Zain et al. [9]. Regarding the effectiveness of Naïve Bayes, SVM weighting in classifying film reviews resulted in an accuracy of 88.8%. Naïve Bayes uses n-gram extraction weighting in its process. In that study, deletion of stop words did not improve classification performance. The author recommends the feature extraction process to use unigram and bigram simultaneously.

Based on previous research, the Naïve Bayes Support Vector Machine results produced better accuracy values than the Support Vector Machine method. So, in this study, we would try to predict the big five personalities using the Naïve Bayes Support Vector Machine method. Naïve Bayes (NB) was a classification method that used probability and statistical methods. Support Vector Machine (SVM) was a classification method that had the convenience of classifying labels using a hyperplane. These two methods had been combined with Naïve Bayes, which would be played a role in weighting. In contrast, the Support Vector Machine would be played a role in classification based on the results of Naïve Bayes weighting. Synthetic Minority Overside Technique (SMOTE) prediction model to handle imbalance data, Bidirectional Encoder from Transformers Representations (BERT) as a semantic approach, and Linguistic Inquiry Word Count as a linguistic feature. To help improve accuracy, the hyperparameter method, namely Optuna, was used. The advantage of using Optuna was that the parameters were built dynamically so that it was more likely to get the best parameters that other hyperparameter methods may not be able to obtain [19]. The dataset used in this study was from previous researched [8].

This paper will be divided into four parts. The first part is an introduction, as described above. The second part is the method used to build the Big 5 Personality prediction system. The third part will explain the results of the experiments, and the last part will present the conclusions of this paper.