CHAPTER 1
INTRODUCTION

This chapter describes the eight subtopics: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework; (4) Problems Statements; (5) Objective; (6) Hypotheses; (7) Scope and Delimitation; and (8) Importance of the Study.

1. 1 Rationale

Twitter is one of the most popular social media in the world. Twitter has active users of approximately 330,000,000 throughout the world in first quarter of 2019 [1]. Twitter has high popularity in a few years. Statistics show that Indonesia has high Twitter usage. This is in line with the previous study which mentioned that Indonesia was ranked as the fifth most tweeting country and 2.4% of worldwide tweets are posted by users in Jakarta [2]. Twitter is often used by users as a tool to publish activities or places to pour out opinions, thought and feelings [3]. Twitter texts, namely tweets which are written in natural language and information from twitter such as the number of followers, following, and mention can be used as a data resource for researchers to gain information. One of the information that can be analyzed from Twitter is user's personality in the form of personality prediction.

Personality prediction is an attempt to find the patterns of behavior, patterns of thinking, and traits of a person towards their environment. Personality prediction system can be used as assessment tools in employee recruitment, relationship counseling and education counseling [4]. One model that can be used to predict personality is the Big Five Personality Model. The Big Five Personality Model has five dimensions that can describe personality, that are Openness, Conscientiousness, Extroversion, Agreeableness, and Emotional Stability [5]. Conscientiousness characterizes people who are careful, dependable, and self-disciplined. Then, Agreeableness characterizes people who are being courteous, good-natured, empathic, and caring. Emotional Stability characterizes people who are poised, secure, and calm. Openness to experience generally refers to the extent

to which people are imaginative, creative, curious, and aesthetically sensitive. Furthermore, Extroversion characterizes people who are outgoing, talkative, sociable, and assertive.

The previous studies in Indonesia show that personality could be determined based on Twitter data using Big Five Personality Model and machine learning models [4] [6] [7]. Based on these studies, the learning algorithm used can be one algorithm such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM) or combination of several algorithms such as Ensemble like Extreme Gradient Boosting (XGB) and Stacking. While, personality prediction system that uses Stacking, has the best performance compared to other classifiers. In addition, these studies applied several techniques such as hyper-parameters optimization, feature selection and sampling techniques that can improve system performance [7]. Furthermore, adding other features which have a correlation with personality besides Twitter texts also affects the performance of system, such as Ong et al. using Twitter user's information as features for Personality Prediction using Big Five Model besides Twitter texts [6].

Grammatical features are one feature that has a correlation with personality [8] [9]. The grammatical features take into information like the number of words used, the amount of punctuation marks used in the text etc. This feature can be taken from text such as twitter texts. This feature has a correlation with personality, especially with Big Five Personality Model [8]. The previous study showed that personality prediction with grammatical features into the Big Five Dimensions using Multilayer Perceptron, Support Vector machines and Naïve Bayes have accurate results in predicting the trait of Extroversion, while the traits of Agreeableness and Emotional Stability also present a high of accuracy [9]. Based on this explanation, grammatical features can be used to improve the performance of a personality prediction system.

Based on these explanation the used of Stacking and adding grammatical features may be the solution to build a personality prediction system that has good performance. The idea of the Stacking algorithm was first introduced by Wolpert [10] in the neural networks context. In the neural networks context there are terms

of hidden layers while in the Stacking there are terms of level learners. Hidden layers in neural networks can be added so that they can affect the system performance. Whereas in the Stacking uniquely focuses on 2 layers or level learners. Furthermore, the second level learner in Stacking work by deducing the biases of the first level learner, so as to produce an optimal model [10]. Stacking has weakness, it performs worse on multi-class than on two-class datasets [11]. While, there is a new approach proposed that is Troika to address multi-class problems [12]. Troika has four level learners. Then, the result showed that Troika performed better than Stacking in terms of classification accuracy. Other studies also show that Stacking with at least one additional level learner can improve performance [13]. Based on several explanations about Stacking show that adding level learners in Stacking can improve the performance of the system, because the aim of adding these level learners is to reduce the bias of the previous level learners so as to produce an optimal model.

1. 2 Theoretical Framework

Grammatical features take into information like the number of words used, the amount of punctuation marks applied in the text etc. These features can be taken from texts that users write. Grammatical features have a correlation with personality, especially with Big Five Personality Model. For example, people who have Openness trait often use question marks, commas, apostrophes, and quotes in the text they write [8]. The used of grammatical features in personality prediction using Big Five Personality Model can be considered to improve system performance.

In the classifier area, Wolpert proposed Stacking algorithm in the neural networks context [10]. The Stacking uniquely focuses on two layers or level learners. In Stacking there are two types of learners called Base Learners and Meta Learner. Base Learners and Meta Learner is the normal machine learning algorithms like Random Forests, SVM, KNN, etc. Furthermore, Meta Learner works by deducing the biases of the Base Learners, so as to produce an optimal model [10]. The Stacking method offers benefits compared with other classifiers,

namely the ability to combine different classifiers with simplicity and having good performance. However, in multi-class dataset, Stacking may perform worse than two-class dataset [11]. Carrying the concept of neural networks that the addition of layers can affect system performance, the addition of level learners on Stacking can also affect performance such as previous study that proposed Troika that containing four level learners [12]. Troika was proposed to address the problem in Stacking, and then the result showed that Troika performed better than Stacking in terms of classification accuracy. Other study also shows that Stacking with at least one additional level learner can improves performance [13].

Related to the performance of multilevel Stacking, this study proposed the use of Stacking by inserting one more level (3rd level learner) in Stacking structure. The proposed method called Modified Stacking. Furthermore, other feature namely grammatical features will be added to improve the personality prediction system performance. This study is different from Troika's approach. It will improve the performance of Stacking and analyze the effects of an additional level learner using only three level learners and grammatical features in personality prediction using Big Five personality model and Indonesian Twitter data.

1. 3 Conceptual Framework

The basic concept of the proposed method is modifying the structure Stacking and adding other features to improve personality prediction system performance. The solution of inserting one more level (3rd level learner) in Stacking structure and adding grammatical features was used for personality prediction. This study is proposed to enrich the studies related to the performance of personality prediction system and multi-level Stacking in classification problems.

1. 4 Problems Statements

Based on theoretical and conceptual framework, the additional features which have a correlation with personality in the prediction system can be

considered to improve system performance, such as adding grammatical features for personality prediction system. While, in the use of classifier for personality prediction, Stacking is a classifier that has the best performance compared to other classifiers. Stacking focuses only on two level learners. It is possible to adding more level learner in the structure of Stacking. Furthermore, the second level learner in Stacking works by deducing the biases of the first level learner, so as to produce an optimal model. Therefore the problem in this study is the additional of level learner in Stacking can affect system performance while only a few studies carried out study in the context of multi-level Stacking, to enrich the study in multi-level Stacking, this study proposed additional a one level learner in Stacking structure and adding grammatical features to improve Stacking performance in personality prediction area.

1. 5 Objective

According to the problem statement, the objectives of this study are to improve the performance of Stacking by modifying the structure Stacking into three level learners and adding grammatical features, so that this approach can be used in personality prediction with high accuracy and better performance than the previous studies. In addition, this study was proposed to enrich to the studies related to the performance of personality prediction system and multi-level Stacking in classification problems.

1. 6 Hypotheses

Improvement of personality prediction systems can be done by using Stacking as a machine learning algorithm and adding features that have a correlation with personality. In the machine learning algorithm area, Stacking is proposed with a structure consisting of two level learners, where the second level learner in Stacking work by deducing the biases of the first level learner, so as to produce an optimal model and improve performance [10]. Then, some studies show that adding level learners in the Stacking can improve performance [12] [13]. Meanwhile, in the additional of feature, grammatical features have a

correlation with personality [8]. The previous study showed that personality predictions with grammatical features have accurate results in predicting personality [9].

Based on the literature review, the hypothesis of this study is that:

The personality prediction process using a modified Stacking method, which consists of three level learners and grammatical features, will produce a more accurate model than Standard Stacking.

The independent variables in this study are grammatical features and level learners of Stacking, and then affect the dependent variable, that is accuracy.

1. 7 Scope and Delimitation

This study formulated the scope and delimitation, they are as follows:

1. This system is designed for personality prediction using Indonesian Twitter dataset. The dataset that used in this study is extracted from the previous study that is Personality Prediction Based on Twitter Information in Bahasa Indonesia [6]. The number of dataset is 250 containing Twitter user information and user's latest tweets which have been labelled with Big Five personality trait.
2. The features that used in this study are user's latest tweets and Twitter user information extracted from previous study [6]. Then, an additional feature is the grammatical features that extracted from previous study [9].
3. Personality Model used as class labels in this study is Big Five personality model. Five classes of Big Five personality are Openness, Conscientiousness, Extroversion, Agreeableness, and Emotional Stability.
4. This study implements some techniques to optimization machine learning that used in classification process, such as Grid Search for hyper-parameter tuning technique, Chi-square for feature selection, and SMOTE sampling technique for handling imbalance of data.
5. This study uses K-NN, Random Forest and Gradient Boosting algorithm as classifier used in Stacking and Modified Stacking method.

6. Data representation uses Bigram to convert sentences into tokens and TF-IDF for feature or tokens weighting.
7. The performance evaluation by the values of accuracy with 10 Fold Cross Validation and dataset ratio is 70:30 (70% training and 30% testing data).

1. 8 Importance of the Study

The contribution of this study is twofold, which are:

1. A modified Stacking method that replaces a two level learner prediction model by a three level learner prediction model to increase the accuracy of personality prediction.
2. The inclusion of grammatical features, in addition to information features, in personality prediction.