JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 267-275
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i1.5439

1. INTRODUCTION

The role of social media in this era has developed, which was initially only for networking, now it is used as a place for self-presentation [1]. Currently, social media is no longer an unfamiliar place for most people to pour out their hearts and minds, as well as a place to voice their opinions. Not surprisingly, social media is a treasure trove for experts to make research material. One example is tweets or retweets shared by users on their social media accounts, which can be used as data for researchers to analyze and predict their personality traits [2].

Personality is a characteristic that reflects how individuals react to their environment. Each individual is unique, and this causes them to have different personalities from each other. This characteristic tends to be hard to change yet not impossible since individuals interact with strength and the environment around them [3]. Many guidelines can be used to predict a person's personality, for example, Big Five Personality Traits, Myers Briggs Type Indicator (MBTI), StrengthsFinder, and DISC Personality [4]. Personality detection using Big Five has been used for a long time to predict how someone's performance is performed in the future, especially for companies recruiting new employees [5]. Predicting someone's performance by looking at the personality highlighted through their social media account is said to provide a better picture when compared to holding a personality test [5].

Several previous studies have been successfully conducted to detect personalities based on their social media data. The previous study by [6] used the Naive Bayes Classifier method to predict Big Five personalities based on Twitter and achieved an accuracy of 42.71%. The dataset used in this study is dominant labeled Agreeableness, with a total of 167 accounts. Meanwhile, the lowest label is Extraversion, with a total of 14 accounts. The author stated that the low accuracy results are caused by the amount of data that is not balanced. A similar study conducted by [7], detected the Big Five personalities via Facebook using the Naive Bayes method. This study was conducted with two test scenarios. The first test consists of a scenario related to data preprocessing whether changing the form of a word will affect the accuracy. Then, the second test consists of a scenario that implements prior probability value. The highest accuracy in this study achieved 59.9% on the first test and 60.2% on the second. Another study conducted by [8], detected the Big Five personalities via Twitter using the Naive Bayes method. The dataset used in this study consisted of 1500 tweet data obtained from 15 accounts, and the psychologist did the labeling directly. Initially, the number of accounts used in this study was 95 accounts. However, to overcome the data imbalance, only three were taken for each Big Five personality type. Thus, the total accounts used are reduced to 15 accounts. The accuracy results obtained in this study were outstanding, which is 86.66%. A previous study conducted by [9], predicted Big Five personalities by comparing several machine learning approaches, consisting of Multinomial Naive Bayes, AdaBoost, and Linear Discriminate Analysis (LDA). Their results stated that Multinomial Naive Bayes performs better than AdaBoost and LDA. Moreover, Multinomial Naive Bayes achieved the highest accuracy score of 73.43% for the feature Openness. A comparative study was also conducted by [10] that evaluated Naive Bayes and Support Vector Machine (SVM) to predict Big Five personality traits. The results showed that Naive Bayes outperformed SVM with an accuracy score of 60%.

Although several previous studies have been successfully conducted, personality detection is still a challenging problem in cognitive computation and a conventional method approach for this problem most likely is inadequate to get promising results [11]. Based on several studies described above, Naive Bayes actually has better performance results for detecting Big Five personality traits than other machine learning methods. However, most of the achieved accuracy score is still not sufficient enough to predict five labels in the Big Five personality. Thus, it is necessary to find out how to optimize machine learning performance for personality detection. Another study that focused on personality detection conducted by [12], predicted Big Five personality traits by combining machine learning and deep learning, which are SVM and IndoBERT. The idea of this research was to implement BERT as a semantic approach and SVM as a classifier. The dataset used in this study consisted of 511,617 tweets from 295 Twitter accounts. As the baseline model, SVM achieved an accuracy score of 57.97%. After combining SVM with BERT, SMOTE, and LIWC, the accuracy score increased to 80.07%. It is shown that combining SVM with BERT can improve the performance result. In previous studies, BERT is still more widely used to handle classification cases such as sentiment analysis and it is still rare to find studies that implement BERT for personality detection [12]. A similar approach by combining machine learning and deep learning was also conducted by [13] to detect emotion from tweets using BERTweet and SVM. The idea of this research was to sum the log probability value from each model. The dataset used in this study has five labels consisting of joy, anger, fear, sadness, and neutral. Their results after combining these two models to predict five labels were outstanding, with an accuracy score of SVM 84%, BERTweet 89%, and the proposed combine

Submitted: **30/12/2022**; Accepted: **19/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 267-275
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i1.5439

model 91%. A previous study conducted by [14] combined BERT with Bayesian Network to classify the governance texts. Same as previous results, this study also performs better by combining machine learning and deep learning with an accuracy score of 94.59%. This accuracy score was 3.23% better than BERT and 18.06% better than Bayesian Network.

Based on the explanation above, it is shown that combining machine learning with deep learning will help the model to improve its performance results. Therefore, this research is aimed to combine Gaussian Naive Bayes and IndoBERT models to detect Big Five personalities based on social media Twitter. According to previous studies, this proposed combined model has the potential to produce a good performance result. This research will experiment to detect Big Five personalities from tweets in Indonesian. It will remove all tweets in another language due to the use of IndoBERT, a BERT base model trained using an Indonesian language document.

Submitted: **30/12/2022**; Accepted: **19/01/2023**; Published: **28/01/2023**