Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. x (2022) x - x ISSN Media Electronic: 2580-0760

Big Five Personality Assessment Using KNN method with RoBERTA

Athirah Rifdha Aryani[1], Erwin Budi Setiawan[2]
[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]athirahrifdha@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

1. Introduction

In this modern era, almost every society has a social network as a means of communication and expressing each user's personal views on different aspects of life. Twitter is a social media that is widely used by several countries to express feelings and activities written in one or two sentences[1]. Internet-based media allow users to interact and express themselves, directly or indirectly, with large audiences or with user-generated content and interactions with others (Caleb T. Carr dan Rebecca A. Hayes, 2015)[2]. Language-based predictions are made by analyzing word choice and word position in a defined category based on the language used. Language analysis was carried out on several social media profiles, the use of everyday language, and short messages[3].

Personality is understood as an individual's state of mind that depends on behavior, emotions, and attitudes such as the differences in the characteristics of each person. The Big Five personality traits are considered an effective way to determine a person's personality because they are more informative.[4]. The Big Five personality traits are often abbreviated as the "OCEAN" model, openness, conscientiousness, extroversion, agreeableness, and neuroticism[1].

Several studies try to measure a person's personality through the Twitter user word classification method. One of them is research conducted[4], The author tries to use the LIWC method to count words automatically based on the category, then use the Support Vector Machine (SVM) method to classify the Big Five Personalities. This study produces a model that can predict a person's personality by 80.07%. The authors show that research can improve the performance of personality prediction systems by collecting more data from respondents and experimenting with different methods such as combining BERT with deep learning to improve the performance of personality prediction systems.

In other research [3] the prediction of the Big Five personality using TF-IDF and with the K-Nearest Neighbour (KNN) method, it can be concluded that the higher accuracy of the components of social behavior and language and by measuring performance in testing the k value = 9 of 60.97%, while the social and linguistic behavior component with a performance of k=1 has a low value with a value of 39.02%.

Other Research [5] conducted research using a semantic approach of the type RoBERTa (Robustly Optimized BERT Approach) Obtaining an accuracy value of 83.2%, the highest value of 81.3%, and the median value of 86.5%.

In this study, the aim is to conduct a test to find a way with a complete formula to obtain a model that can improve the classification accuracy and personality prediction of the Big Five using K-Nearest Neighbours. To improve the accuracy achieved in K-Nearest Neighbours by combining LIWC, Information Gain, RoBERTa, K-Means SMOTE, and hyperparameter tuning Tested.

We build a personality detection to predict a person's 5 big personality traits using the K-Nearest Neighbours (KNN) method. The K-Nearest Neighbours method classifies the subjects based on the training data closest to the subject to be used as the Big Five Personality classifier. K-Means SMOTE (Synthetic Minority Synthetic Engineering) is a predictive model for dealing with imbalanced data, Information Gain is used as a feature selection method with the highest feature rating is the most relevant feature and has closely related to the linked dataset, Robustly BERT Approach (RoBERTA) as semantic approach and Linguistic Inquiry Word Count as linguistic feature word counts can improve performance based on correlations

Accepted: xx-xx-2022 | Received in revised: xx-04-2022 | Published: xx-04-2022

between speech and psychologically relevant text. Hyperparameter settings will be added to find the best setting for KNN and help improve accuracy.

This research is structured as follows. Section 2 describes how Twitter's personality prediction system was investigated. Section 3 presents the results of the experiment and discussion and Section present conclusions.

DOI: https://doi.org/10.29207/resti.v6iX.xxx