PREDICTING GEN-Z PERSONALITY ON TWITTER BASED ON BIG FIVE MODEL WITH KNN AND SVM

Aang Kisnu Darmawan[1]; Salman Alfarisi[2*]; Hozairi[3]

Department of Information System[1, 2], Department of Informatics[3]
Universitas Islam Madura, Pamekasan, Indonesia[1, 2, 3]
www.uim.ac.id[1, 2, 3]
ak.darmawan@gmail.com[1], salman.alfarisi52002@gmail.com[2*], dr.hozairi@gmail.com[3]
(*) Corresponding Author

Abstract— *Generation Z is a group that is very connected to digital technology, especially social media such as Twitter. Their widespread presence on these platforms creates a unique opportunity to understand their behavioural patterns and personalities. However, research on personality prediction on social media is still limited and focused on certain platforms or different age groups. Personality prediction can help to find out someone's personality by just looking at tweets on social media. This research aims at two things: first, to build a Gen-Z personality prediction model on Twitter based on the Big Five Personality Model with the K-Nearest Neighbor (KNN) algorithm and Support Vector Machine (SVM). Second, test and compare the performance of previously generated personality prediction models with various evaluation metrics. The research results show that the KNN algorithm has an accuracy rate of 0.73%, precision of 0.73%, recall of 0.73%, and score of 0.72%. Based on the test results, the SVM algorithm obtained the best accuracy, which received an accuracy of 0.78%, precision of 0.82%, recall of 0.78%, and F1-score of 0.78%. This research contributes in two ways: first, scientifically, by understanding Gen-Z personalities on Twitter, and second, by developing new prediction methods and insights into Gen-Z behaviour. Second, practically, by helping with communication and marketing strategies, product/service development and social interventions for Gen-Z.*

Keywords: *machine learning, prediction algorithm, big five personality, KNN, SVM*

Abstrak— *Generasi Z merupakan kelompok yang sangat terhubung dengan teknologi digital, terutama media sosial seperti Twitter. Kehadiran mereka secara luas di platform ini menciptakan kesempatan unik untuk memahami pola perilaku dan kepribadian mereka. Namun, Sejauh ini, penelitian tentang prediksi kepribadian di media sosial masih terbatas dan terfokus pada platform-platform tertentu atau kelompok usia yang berbeda prediksi ke pribadian dapat membantu untuk mengetahui kepribadian seseorang dengan hanya melihat tweet di media sosial. Penelitian ini bertujuan dua hal: pertama, membangun model prediksi kepribadian Gen-Z di Twitter berdasarkan Model Personality Big five dengan algoritma K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM). Kedua, menguji dan membandingkan performa model prediksi kepribadian yang dihasilkan sebelumnya dengan berbagai metrik evaluasi. Hasil penelitian menunjukkan bahwa algoritma KNN memiliki tingkat accuracy 0,73%, precision 0,73%, recall 0,73% , F1score 0,72%. dan algoritma SVM memperoleh hasil akurasi terbaik diperoleh oleh berdsarkan hasi pengujian didapatkan accuracy 0,78%, precision 0,82%, recall 0,78% dan F1-score 0,78%. Penelitian ini memberikan kontribusi dalam dua hal: pertama, secara saintifik dengan mengembangkan pemahaman tentang kepribadian Gen-Z di Twitter, metode prediksi baru dan wawasan perilaku Gen-Z. Kedua, secara praktis dengan membantu strategi komunikasi dan pemasaran, pengembangan produk/layanan dan intervensi sosial untuk Gen-Z.*

Kata Kunci : *machine learning, algoritma prediksi, bigfive personality, KNN, SVM*

INTRODUCTION

Generation Z is a group that is very connected to digital technology, especially social media such as Twitter (Pujiono, 2021). The Gen-Z population of social media users in Indonesia is very large and influential. With approximately 65.6 million active social media users, Gen-Z is an important segment for various industries,

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 2 September 2024 | DOI: 10.33480/pilar.v20i2.5580

including digital marketing, entertainment, and education (Firamadhina & Krisnani, 2021). Their widespread presence on these platforms creates a unique opportunity to understand their behavioural patterns and personalities (Trang et al., 2024). Research on Generation Z personality predictions on Twitter social media is urgent because it can provide valuable insight into how technology and social media shape the identity and characteristics of this generation (Utami et al., 2021).

Recent research has focused on predicting Gen Z personalities using machine learning techniques. Research has shown that personality prediction models can be developed effectively by incorporating domain knowledge, such as lexicons related to psycholinguistics and mental health, to improve accuracy and interpretability (Maharani & Effendy, 2022). Machine learning algorithms, such as SVM, have been used to predict personality traits based on social media data, showing increased prediction accuracy (Han et al., 2023). Additionally, the use of machine learning in classifying individuals based on their personality features has been highlighted, with applications in marketing campaigns and competitive exams for talent evaluation (Gupta et al., 2023). Additionally, innovative approaches, such as the "Predictive Five" model, have been proposed to provide a more accurate and interpretable representation of human personality, showing promising results in predictive performance (Wang et al., 2022).

Additionally, logistic regression models for Big-Five personality assessment from a series of questions have been proposed as a cost-effective, real-time solution in the Big Five Personality model, there are five main dimensions used to describe an individual's personality. First, Openness reflects how open a person is to new experiences, creative ideas, and abstract thinking. Second, Conscientiousness measures how organized, disciplined, and self-controlled a person is. Third, Extraversion indicates the level of activity, energy, and tendency to engage in social interactions. Fourth, Agreeableness describes how cooperative, friendly, and empathetic a person is towards others. Lastly, Neuroticism reflects a person's susceptibility to negative emotions such as anxiety or depression (Maulidah, 2023). Moreover, utilising the k-means clustering algorithm in categorising personalities according to the Big Five model has been emphasised as a beneficial resource for organisations in selecting candidates and providing career assistance (Vijay & Sebastian, 2022). Leveraging social media data for personality prediction through methods such as Linear Discrimination Analysis, Multinomial Naive Bayes, and AdaBoost shows the potential of utilising textual and image data for accurate personality assessment (Cai & Liu, 2022). These diverse approaches collectively contribute to advancing the understanding and prediction of Gen-Z personality through machine learning methodologies.

Research into Gen-Z personality predictions on Twitter faces several challenges. First, a lack of understanding of Gen-Z's unique behavioral patterns on social media platforms such as Twitter hinders accurate personality predictions (Gade et al., 2023). Second, Twitter data's complexity and unstructured nature pose challenges in achieving high prediction accuracy, highlighting the need for sophisticated machine-learning algorithms (Tandon & Mehra, 2023). Third, previous studies may have used non-specific data or suboptimal algorithms for Twitter, limiting the effectiveness of personality prediction models (Han et al., 2023). Lastly, there is a gap in exploring the practical implications of Gen-Z personality prediction research on Twitter, with findings not widely applied in real-world scenarios, indicating the need for further research to bridge this gap (Mustafa et al., 2023).

Previously, several studies on personality prediction only utilized personality data obtained from the Kaggle Machine Learning Repository. This data was used as training and testing data for personality prediction models based on the Big Five Personalities (Aisah et al., 2022). Several studies have used data from platforms like Facebook or Instagram, but little research has explored Twitter specifically. This research aims to contribute to the literature by looking at the relationship between behaviour on Twitter and the dimensions of the Big Five Personality in Generation Z. Various previous studies have been conducted (Xu et al., 2021).

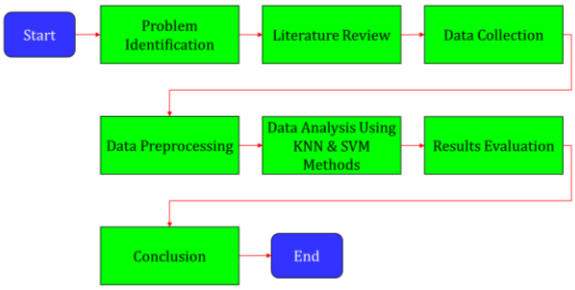
This research aims to achieve two objectives: first, to construct a personality prediction model for Gen-Z on Twitter based on the Big Five Personality Model using the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms this study selects the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms to build the personality prediction model for Gen-Z on Twitter. KNN is chosen for its flexibility in handling various types of data without assuming specific data distributions. SVM is selected for its capability to manage high-dimensional and complex data, and its ability to find the optimal separator between data classes with the largest margin using the kernel trick. Both algorithms have proven effective in numerous cases, including text analysis and data-driven social media predictions such as Twitter. Secondly, this research tests and

compares the performance of the previously generated personality prediction models using various evaluation metrics.

This research makes a valuable contribution to understanding and effectively reaching Gen-Z. Scientifically, this research expands knowledge about how "The Big Five" model can be applied to understand personality on social media, especially in the context of Gen-Z on Twitter. The research results can help marketers and communicators better understand the Gen-Z target audience and develop more effective communication and marketing strategies. Developing products and services that suit the needs and preferences of Gen-Z can also be done by utilising the results of this research. The findings of this research can also be used to design more effective social interventions to reach Gen-Z and address the social problems facing this generation.

MATERIALS AND METHODS

The data analysis method used is the K-Nearest Neighbors Method and Support Vector Machine. This method is one of the methods available in classification techniques in Data Mining. The research stages are as shown in the following picture:



Source: (Research Results, 2024)
Figure 1. Research Stages

Below is an explanation of the stages according to the diagram in Figure 1 above.

A. Problem Identification

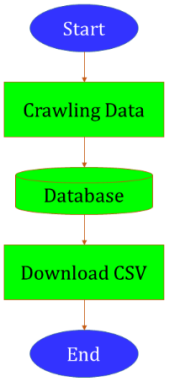
According to various personality prediction studies. Social media personality prediction study is limited to select platforms and age groups. Several studies have used Facebook or Instagram data, but few have examined Twitter. This study examines Generation Z's Twitter behaviour and Big Five Personalities to add to the literature. With various personality theories, the Big Five is weighted using Term Frequency Inverse Document Frequency. Data from data crawling used K-Nearest Neighbour and Support Vector Machine for classification since it is fast, simple, and accurate.

B. Literature Review

A literature review is a systematic examination and analysis conducted to comprehend the research and works previously conducted on a specific topic. The objective of a literature study is to gather, assess, and integrate information published in scientific literature, journals, books, and other pertinent sources related to a specific research or study subject.

C. Data Collection

Currently, data is gathered via questionnaires to randomly selected respondents who are willing to participate and have an accessible Twitter account for one month. articipants were requested to respond to enquiries using the globally recognised IPIP (International Personality Item Pool) standard The reference for the questionnaire can be accessed for free at the following link: https://bigfive-test.com/. In the next stage, collection was conducted using a tweet harvesting tool developed with Python on Google Colab. The process involved retrieving up to 100 of the most recent tweets from each respondent, which were subsequently categorized according to the Big Five personality traits. The dataset was acquired from Twitter in 2024 and the data collection period spanned approximately one month, from May 1 to May 31. To ensure the data's relevance and accuracy, only the most recent tweets were crawled, and the focus was solely on tweets in Indonesian. The collected data was then formatted and stored in CSV files to facilitate the prediction of personality traits.



Source: (Research Results, 2024)
Figure 2. Crawling Data

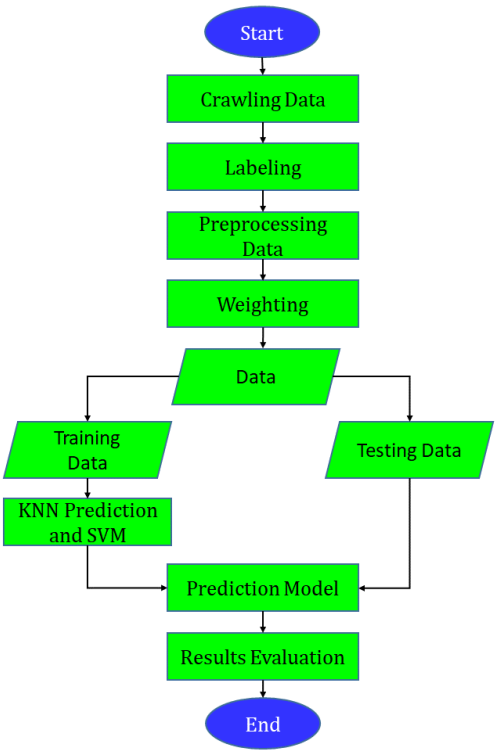
D. Data Pre-processing

Before performing classification, a pre-processing step is conducted to enhance the effectiveness and efficiency of the mining process. During this stage, the process involves collecting data, cleaning it, tokenising it, converting it to lowercase, removing stopwords, and performing stemming.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 2 September 2024 | DOI: 10.33480/pilar.v20i2.5580

E. Data Analysis with KNN and SVM

At this stage of analysing the results, the author attempts to conduct a data analysis process using Python to predict Gen-Z personalities on Twitter social media with the Big Five Personality prediction model. After that, a labelling process is carried out based on the results of personality tests, which are distributed voluntarily and involve psychology experts. This labelling process is then weighted using the TF-IDF weighting method. After the number weighting stage, the next classification process uses the K-Nearest Neighbor Support algorithm and Support Vector Machine to predict data from the labelling done through the previous weighting results. This process uses the Python programming language. So that accurate results will be obtained. This accuracy value is the percentage accuracy of a data record that will be classified correctly after testing the classification results.



Source: (Research Results, 2024)

Figure 3. Stages of the KNN and SVM methods

The stages in the data analysis process are depicted in Figure 3 above. The approach involves applying the TF-IDF method for weighting. The data processing stage is the initial step of crawling data from the social media platform Twitter. Subsequently, the process of labelling is executed. The collected questionnaire data have been categorised into five Big Five personality traits. This labelling method involves examining individuals using the Big Five Inventory (BFI) and relies on the expertise of psychology professionals to provide labels. The objective of this step is to determine the personality type and Twitter account name of each participant. The subsequent stage involves cleaning and preparing the data through several steps: cleansing, case folding, normalisation, stopword removal, and tokenising. These steps are then followed by a classification process using the K-Nearest Neighbour Support Vector Machine method. Subsequently, once the classification has been completed, the metrics testing procedure is conducted, along with the categorisation of the report. This will unveil the accuracy, precision, recall, and f1 score values.

F. Result Evaluation

Once the data has been classified, the subsequent step involves assessing the accuracy, recall, and precision to gauge the algorithm's effectiveness in classifying the data. An evaluation is conducted to examine the performance of the model. Testing assesses the system's efficacy in evaluating the data employed. This test utilises the Python library, namely the classification_report function from sci-kit-learn. The classification_report function is a tool used to evaluate the performance of a classification model. It summarises several important metrics: accuracy, precision, recall, and f1_score. This report functions by juxtaposing the prognostications generated by the model with the factual data.

RESULTS AND DISCUSSION

In this section, we will discuss the results of data analysis based on data obtained from Twitter social media, which has been crawled using Python on Google Colab, which contains a collection of Twitter on Twitter social media to predict their personality based on the Big Five Personality prediction model.

A. Data Collection Results

At this stage, data scraping or collection was conducted using Python libraries in Google Colab. The data collected for this study originated from Twitter users who voluntarily filled out the questionnaire. Responses varied regarding the distribution of questionnaires; some agreed to participate while others declined. Ultimately, 120 respondents agreed to participate in this study. For each respondent, a maximum of their 100 most recent tweets were collected. During the data collection phase, the author gathered the most recent tweets to ensure the relevance and timeliness of the data used in the analysis. The collected Twitter data included a total of 1.200 tweets, with 100 most recent tweets from each

respondent's account. The collected data consisted of tweet dates (pupdate), authors (usernames), and texts (tweet contents). Pupdate refers to the timestamp of the tweet creation, author indicates the username of the tweet creator, and text represents the content originating from the tweet that has been created. Once collected, these tweets were transformed into a structured table format to facilitate further processing. For a clearer overview of the data quantity obtained, please refer to Table 1 below.

Table 1. Data Crawling Results

Media social	Crawling Result
Tweet	1200

Source: (Research Results, 2024)

The next stage is the labelling process, where the collected questionnaire results are labelled into five Big Five character categories. This process is based on the BFI (Big Five Inventory) assessment and involves psychology experts conducting the labelling. This step was carried out to determine each respondent's personality label type.

B. Data Pre-Processing Results

a. Cleaning & Case Folding Results

At this stage, the existing data is raw data that has not been processed. Several stages are carried out: deleting symbols, deleting URL links (https and http), removing punctuation marks, considering letters and numbers, changing each word to lowercase, and separating and combining data. Then, at this stage, duplicate and empty data in tweets is removed. The following are the results of the cleaning and case-folding process:

Table 2. Cleaning & Case Folding Results

Description	Clean
Menjelajahi beragam topik dari teknologi hingg...	menjelajahi beragam topik dari teknologi hingg...
Setiap hari adalah peluang baru untuk bertemu ...	setiap hari adalah peluang baru untuk bertemu ...
Setiap hari dimulai dengan jadwal yang sudah d...	setiap hari dimulai dengan jadwal yang sudah d...

Source: (Research Results, 2024)

From Table 2 above, it is known that there are several changes in the data results before they are processed with data that has undergone a cleaning process. At this cleaning stage, you can also change the amount of existing data by removing columns and duplicates.

b. Tokenizing Results

At this stage, we use the Python library, namely NLTK, to break down a sentence in the text into chunks of words, which can then be processed easily. The following are the results of the tokenisation process.

Table 3. Tokenising Results

Clean	Token
menjelajahi beragam topik dari teknologi hingg...	[menjelajahi, beragam, topik, dari, teknologi,...
setiap hari adalah peluang baru untuk bertemu ...	[setiap, hari, adalah, peluang, baru, untuk, b...
setiap hari dimulai dengan jadwal yang sudah d...	[setiap, hari, dimulai, dengan, jadwal, yang, ...

Source: (Research Results, 2024)

Tokenising, which is carried out in Table 3 above, is the stage of breaking down sentences into chunks of words such as “setiap hari adalah peluang baru untuk bertemu” into “setiap, hari, adalah, peluang, baru, untuk, bertemu” This is done to make it easier to weight the scores in the process weighing.

c. Stopword Removal Result

This process is carried out to remove unimportant words, or stop words, from Indonesian text. The first step in this process is to import the NLTK (Natural et al.) module and download the Indonesian stopword list from NLTK. Then, the Indonesian stopword list from the CSV file that was previously created manually is loaded and saved in CSV format. The list of stopwords in the CSV file is converted into sets, and some special stopwords such as "yg", "utk", "tdk", or some words that have no meaning like "adalah", "tidk", "jangan", and so on are also added to the "stopwords" variable. The following is an example of Stopword Removal results.

Table 4. Results of Stopwords Removal

Token	Stopword
menjelajahi, beragam, topik, dari, teknologi,...	[menjelajahi, beragam, topik, teknologi, seni,....
[setiap, hari, adalah, peluang, baru, untuk, b...	peluang, bertemu, orang, orang, menjalin, hub...
[setiap, hari, dimulai, dengan, jadwal, yang, ...	[jadwal, direncanakan, disiplin, kunci, dailyr..

Source: (Research Results, 2024)

Stopwords removal carried out in Table 4 above is the stage of deleting unimportant words such as in the sentence “setiap, hari, dimulai, dengan, jadwal, yang” becomes “jadwal, direncanakan, disiplin, kunci,”the words deleted are “setiap, hari, dimulai” this was removed because it is not important.

d. Stemming Results

In this process, a preprocessing stemming stage is carried out by utilising the stemmer factory library in the Python programming language. To make data processing easier. This

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 2 September 2024 | DOI: 10.33480/pilar.v20i2.5580

stemming stage aims to be able to remove affixes and replace words with their basic word forms. The following are the results of the stemming process.

Table 5. Stemming Results

Stopword	Stemming
[menjelajahi, beragam, topik, teknologi, seni,....	[jajah, agam, topik, teknologi, seni, percaya,...
[peluang, bertemu, orang, orang, menjalin, hub...	[peluang, temu, orang, orang, jalin, hubung, h...
[jadwal, direncanakan, disiplin, kunci, dailyr..	[jadwal, rencana, disiplin, kunci, dailyroutine

Source: (Research Results, 2024)

In Table 5 above is the stemming stage. This stage aims to remove affixes in each word, such as the word “direncanakan” in the first table, which is changed to "rencana” due to the stemming process.

C. TF-IDF Weighting

Before implementing the KNN and SVM algorithms, data that has completed the preprocessing stage will be weighted using the TF-IDF method. This method calculates the Term Frequency (TF) and Inverse Document Frequency (IDF) values for each token (word) in each document in the corpus.

Table 6. TF-IDF Weighting Results

No	Abadi	Abai	Abe	Abhipraya
0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0

Source: (Research Results, 2024)

In Table 6, this is the stage of weighting the text in the "deskripsi” column into a vector representation using TF-IDF, which gives weight to words based on their frequency of appearance in the document and the entire document collection. The data becomes ready for further analysis by representing text in vector form.

D. K-Nearest Neighbor Classification

At this stage, the algorithm used is K-Nearest Neighbor (KNN). The researcher divided the data into training and testing data with an 80:20 ratio. Probability calculations were performed using Python on Google Colab, and the results obtained can be seen in Figure 4 below.

After the classification stage is carried out, the performance test of the KNN algorithm will continue. This is to determine how effective the KNN algorithm is in making personality predictions. The test method used in performance testing is the matrix test. Based on the test results, the accuracy was 0.73%, precision 0.73%, recall 0.73%, score 0.72%.

```
Accuracy: 0.7302904564315352
Precision: 0.7320450891711188
Recall: 0.7302904564315352
F1-score: 0.728064855672104

Classification Report:
                  precision    recall  f1-score   support

          Openness       0.67      0.69      0.68        42
Conscientiousness        0.72      0.86      0.79        66
      Extroversion       0.72      0.62      0.67        47
     Agreeableness       0.77      0.72      0.75        57
       Neuroticism       0.77      0.69      0.73        29

          accuracy                           0.73       241
         macro avg       0.73      0.72      0.72       241
      weighted avg       0.73      0.73      0.73       241
```

Source: (Research Results, 2024)

Figure 4. Accuracy of the SVM algorithm

E. Support Vector Machine Classification

In the classification stage, the algorithm used is Support Vector Machine (SVM). The researcher divided the data into training and testing datasets with an 80:20 ratio. Probability calculations were conducted using Python in Google Colab. From the calculation results, an accuracy of 78% was obtained, as shown in Figure 5 below.

```
Accuracy: 0.7883817427385892
Precision: 0.8275987309031021
Recall: 0.7883817427385892
F1-score: 0.7839921321938957

Classification Report:
                  precision    recall  f1-score   support

          Openness       0.90      0.67      0.77        42
Conscientiousness        0.68      0.95      0.80        66
      Extroversion       1.00      0.57      0.73        47
     Agreeableness       0.75      0.91      0.83        57
       Neuroticism       0.91      0.69      0.78        29

          accuracy                           0.79       241
         macro avg       0.85      0.76      0.78       241
      weighted avg       0.83      0.79      0.78       241
```

Source: (Research Results, 2024)

Figure 5. KNN Algorithm Accuracy

After the classification stage is carried out, the performance test of the SVM algorithm will be continued. This is to determine how effective the SVM algorithm is in making personality predictions. The test method used in performance testing is the matrix test. Based on the test results, the accuracy was 0.78%, precision 0.82%, recall 0.78%, and score 0.78%.

F. Comparison of Accuracy Results

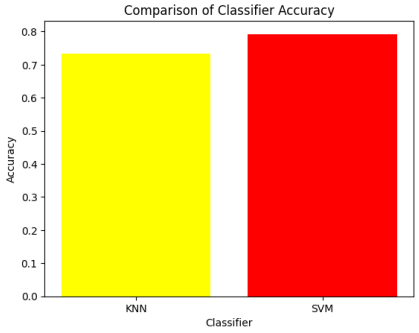
A comparison of the accuracy results of the KNN method (0.73) and SVM (0.78) shows that the SVM method has higher accuracy for predicting Gen-Z personalities on Twitter social media. The comparison results of the two algorithms can be seen in the following graph.

The K-Nearest Neighbors (KNN) method achieves an accuracy of 73%. This method works by classifying data based on its proximity to other data whose class is already known. Although quite effective, KNN has several disadvantages,

especially in high computational requirements for large data and dependence on the selection of 'k' parameters.

On the other hand, the Support Vector Machine (SVM) method achieves higher accuracy, namely 78%. SVM works by finding the optimal hyperplane that separates data from different classes by the largest margin. The advantage of SVM lies in its ability to handle high-dimensional data and its reliability in situations where data is not linearly separated through kernel tricks.

The graph in Figure 6 below displays a visual comparison of the accuracy of both methods, demonstrating SVM's superiority over KNN in predicting Gen-Z personalities on the social media platform Twitter.

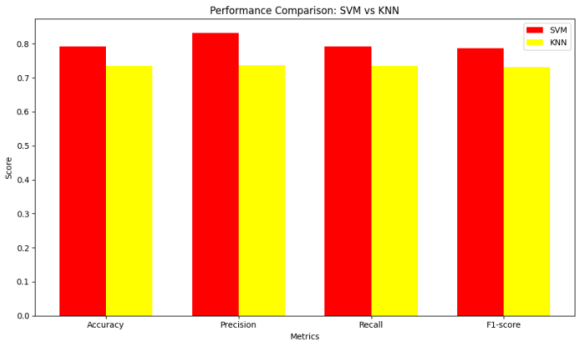


Source: (Research Results, 2024)

Figure 6. Comparison of Accuracy Classification

G. Performance Comparison

After obtaining the accuracy results from KNN and SVM, the next step is to compare the results of both methods. It was found that the SVM algorithm has higher accuracy, precision, recall, and F1 score compared to KNN. The comparison results of the two algorithms can be seen in the graph in Figure 7 below.



Source: (Research Results, 2024)

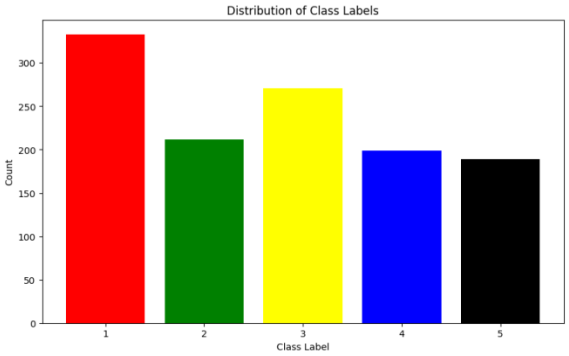
Figure 7. Comparative Performance of KNN and SVM

H. Graphic Label Distribution Graph

It can be seen that the highest class of predicted results is the Openness class (1) with several 350. This shows that most individuals in the sample analysed tend to be open to new experiences, creative ideas and abstract thinking. These characteristics reflect curiosity, a strong imagination, and an appreciation of art and aesthetics.

The Conscientiousness class (2) has a class distribution of 210. Individuals in this category tend to have high discipline, are responsible, and are very organised. They are known to be thorough in planning and carrying out their tasks diligently and have a strong drive to achieve their goals. The Extroversion class (3) is the second largest class, with a class distribution 280. This shows that many individuals in the sample have extrovert traits, generally characterised by social, energetic and assertive traits. They tend to enjoy social interactions, like being in crowds and having a high activity level. The Agreeableness class (4) has a class distribution of 200. People in this category usually show friendly, cooperative and empathetic traits. They tend to put other people's interests first, have a warm attitude, and do not like conflict.

The lowest class, on the other hand, is the Neuroticism class (5) with a count of 180. This indicates that fewer individuals in the sample exhibit neurotic tendencies. Individuals with high neuroticism scores typically experience more anxiety, are more susceptible to stress, and exhibit higher emotional fluctuations. Conversely, low scores in this dimension indicate emotional stability and stress resilience. The distribution chart of labels can be seen in Figure 8 below.



Source: (Research Results, 2024)

Figure 8. Label Distribution

I. Discussion

This research was conducted because Generation Z is the largest group of digital technology users, especially on social media such as Twitter. The Gen-Z population of social media users in Indonesia is both large and influential. Their extensive presence on these platforms provides a unique opportunity to understand their behavioral patterns and personalities. To date,

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 2 September 2024 | DOI: 10.33480/pilar.v20i2.5580

research on personality prediction through social media remains limited, often focusing on specific platforms or age groups. While several studies have utilized data from platforms such as Facebook or Instagram, there has been relatively little research exploring Twitter, especially in relation to predicting Generation Z personalities. Twitter is particularly important as it offers valuable insights into how technology and social media influence the identity and characteristics of this generation. This research aims to develop an accurate and valid prediction model while also providing valuable insights and practical methods to help various stakeholders better understand and engage with Gen-Z. Data was collected by extracting Twitter data through a data crawling process using Twitter's API and Python's Tweet Harvest tool on Google Colab.

In this study, to determine Gen-Z personalities based on the Big Five model using SVM and KNN, we identified significant words or features from relevant tweet text analysis. Each dimension of the Big Five personality traits has its own characteristics. For instance, for Openness, we selected words reflecting creativity, interest in culture, and curiosity about new experiences. Examples include words such as "creative," "innovative," and "exploration," which serve as relevant indicators. For Conscientiousness, we focused on words indicating responsibility, orderliness, and diligence in tasks. Words like "organized," "responsible," and "time management" reflect the pertinent traits in this dimension. We employed a similar approach for other dimensions such as Extraversion, Agreeableness, and Neuroticism, selecting words that reflect levels of social energy, empathy, and responses to stress respectively. By considering these words in our analysis, we identified characteristics relevant to each Big Five personality dimension, supporting the construction of an accurate and valid prediction model for the Gen-Z population on the Twitter platform.

This research shows that classification using K-Nearest Neighbor machine learning produces an accuracy value of 73%, precision of 73%, recall of 73%, and F1-score of 72%. Meanwhile, the Support Vector Machine algorithm provides the best results with an accuracy value of 78%, precision of 82%, recall of 78%, and F1-score of 78% In this study, SVM achieves higher accuracy compared to KNN due to its superior approach in finding the separating boundary between various Gen-Z personalities on Twitter. SVM utilizes complex mathematical techniques to handle intricate data and adapts well to nonlinear patterns. The use of kernel trick and parameter optimization such as C and choice of kernel aid SVM in enhancing prediction accuracy. Besides its technical characteristics, it's crucial to check the balance of label data to ensure accurate interpretation of results. In this regard, testing should account for a balanced distribution of Big Five personality labels (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) in the dataset to ensure validity and generalizability of the constructed model. The highest class distribution from the prediction results is the Openness class, with a total of 350, while the lowest class is the Neuroticism class, with a total of 180.

In previous research, research on personality (Ichsanudin et al., 2021) prediction on social media was carried out by Personality Prediction Based on Twitter Social Media Using the Naïve Bayes Classification Method. The method used in this personality prediction research was carried out to classify a tweet into five personalities. The personality method researchers use is the Big Five Personality, which consists of openness, conscientiousness, extraversion, agreeableness, and neuroticis with classification calculations using NaiveBayes. The result of this research is an accuracy of 42% with the largest class, namely Agreeableness. Meanwhile, this research collects data from Twitter social media and processes it using the K-Nearest Neighbor algorithm and Support Vector Machine. The Term Frequency (TF) and Inverse Document Frequency (IDF) methods are used for data weighting. The programming language used is Python, with a division of training data and test data of 80:20 for each set of data tested.

CONCLUSION

This research aims at two things: first, to build a Gen-Z personality prediction model on Twitter based on the Big Five Personality Model with the K-Nearest Neighbor (KNN) algorithm and Support Vector Machine (SVM). Second, test and compare the performance of previously generated personality prediction models with various evaluation metrics. Based on test results, KNN has an accuracy of 73%, precision of 73%, recall of 73%, and F1-score of 72%. Meanwhile, SVM gave the best accuracy results with 78% accuracy, 82% precision, 78% recall and 78% F1-score. The highest class distribution of prediction results is the Openness class, with a total of 350, while the lowest class is the Neuroticism class, with a total of 180. A comparison of the accuracy results between KNN (73%) and SVM (78%) shows that the SVM method has higher accuracy in predicting Gen-Z personalities on Twitter social media. However

this research still has shortcomings due to limited data. To enhance the quality of this research, it is advised to increase the amount of data used in the prediction process and explore other machine learning methods. Additionally, incorporating manual labeling with input from psychology and language experts could help compare results and improve performance, ensuring that the findings are better aligned with real-world contexts.

REFERENCE

Aisah, S., Umbara, F. R., & Ashaury, H. (2022). Klasifikasi Kerpibadian Berdasarkan Big Five Personality Menggunakan Metode Fuzzy Decision Tree Dengan Algoritma C4.5. *Jurnal Teknologi Informatika Dan Komputer*, *8*(1), 333–349. https://doi.org/10.37012/jtik.v8i1.1110

Cai, L., & Liu, X. (2022). Identifying Big Five personality traits based on facial behavior analysis. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1001828

Firamadhina, F. I. R., & Krisnani, H. (2021). Perilaku Generasi Z Terhadap Penggunaan Media Sosial Tiktok: TikTok Sebagai Media Edukasi dan Aktivisme. *Share : Social Work Journal*, *10*(2), 199. https://doi.org/10.24198/share.v10i2.31443

Gade, P. D. M. V., Kharat, A., Thoke, S., & Khare, K. (2023). Review on Text-Based Personality Prediction Using Social Media Data. *International Journal of Innovations in Engineering and Science*, *8*(4). https://doi.org/10.46335/IJIES.2023.8.4.4

Gupta, I., Jain, M., & Johri, P. (2023). Smart-Hire Personality Prediction Using ML. *2023 International Conference on Disruptive Technologies (ICDT)*, 381–385. https://doi.org/10.1109/ICDT57929.2023.10151367

Han, N., Li, S., Huang, F., Wen, Y., Su, Y., Li, L., Liu, X., & Zhu, T. (2023). How social media expression can reveal personality. *Frontiers in Psychiatry*, *14*. https://doi.org/10.3389/fpsyt.2023.1052844

Ichsanudin, M., Susilo, A., & Solehudin, A. (2021). Prediksi Kepribadian Berdasarkan Media Sosial Twitter Menggunakan Metode Naïve Bayes Classifier. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 988–996. https://doi.org/http://dx.doi.org/10.30645/j-sakti.v5i2.394

Maharani, W., & Effendy, V. (2022). Big five personality prediction based in Indonesian tweets using machine learning methods. *International Journal of Electrical and Computer Engineering (IJECE)*, *12*(2), 1973. https://doi.org/10.11591/ijece.v12i2.pp1973-1981

Maulidah, M. (2023). Klasifikasi Kepribadian Menggunakan Algoritma Machine Learning. *Jurnal Informatika Dan Tekonologi Komputer (JITEK)*, *3*(1), 66–73. https://doi.org/10.55606/jitek.v3i1.1292

Mustafa, H., Mansoor, T., Yaqoob, S., Mehmood, S., & Rahman, M. M. (2023). Personality Analysis by Tweet Mining. *Journal of Innovative Computing and Emerging Technologies*, *3*(1). https://doi.org/10.56536/jicet.v3i1.60

Pujiono, A. (2021). Media Sosial Sebagai Media Pembelajaran Bagi Generasi Z. *Didache: Journal of Christian Education*, *2*(1), 1. https://doi.org/10.46445/djce.v2i1.396

Tandon, V., & Mehra, R. (2023). Study of Approaches to Predict Personality Using Digital Twin. In *Neoromomorphic computing* (p. 296). https://doi.org/10.5772/intechopen.110487

Trang, N. M., McKenna, B., Cai, W., & Morrison, A. M. (2024). I do not want to be perfect: investigating generation Z students' personal brands on social media for job seeking. *Information Technology & People*, *37*(2), 793–814. https://doi.org/10.1108/ITP-08-2022-0602

Utami, N. A., Maharani, W., & Atastina, I. (2021). Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method. *Procedia Computer Science*, *179*, 177–184. https://doi.org/10.1016/j.procs.2020.12.023

Vijay, H., & Sebastian, N. (2022). Personality Prediction using Machine Learning. *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–6. https://doi.org/10.1109/IC3SIS54991.2022.9885425

Wang, B., Shen, C., Zhao, T., Zhai, X., Ding, M., Dai, L., Gai, S., & Liu, D. (2022). Development of a Check-All-That-Apply (CATA) Ballot and Machine Learning for Generation Z Consumers for Innovative Traditional Food. *Foods*, *11*(16), 2409. https://doi.org/10.3390/foods11162409

Xu, J., Tian, W., Lv, G., Liu, S., & Fan, Y. (2021). Prediction of the Big Five Personality Traits Using Static Facial Images of College Students With Different Academic Backgrounds. *IEEE Access*, *9*, 76822–76832. https://doi.org/10.1109/ACCESS.2021.3076989