Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 4 (2021) 680 - 687 ISSN Media Elektronik: 2580-0760

# Semantic Approach for Big Five Personality Prediction on Twitter

Ghina Dwi Salsabila[1], Erwin Budi Setiawan[2]
[1,2]Informatics, School of Computing, Telkom University
[1]ghinghina@student.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

## Abstract

*Personality provides a deep insight of someone and has an important part in someone's job performance. Predicting personality through social media has been studied on several research. The problem is how to improve the performance of personality prediction system. The purpose of this research is to predict personality on Twitter users and increase the performance of the personality prediction system. An online survey using Big Five Inventory (BFI) questionnaire has been distributed and gathered 295 Twitter users with 511,617 tweets data. In this research, we experiment on two different methods using Support Vector Machine (SVM), and the combination of SVM and BERT as the semantic approach. This research also implements Linguistic Inquiry Word Count (LIWC) as the linguistic feature for personality prediction system. The results showed that combination of these two methods achieve 79.35% accuracy score and with the implementation of LIWC can improve the accuracy score up to 80.07%. Overall, these results showed that the combination of SVM and BERT as the semantic approach with the implementation of LIWC is recommended to gain a better performance for the personality prediction system.*

*Keywords: Big Five Personality, BERT, SVM, LIWC*

## 1. Introduction

Personality is well-known as a mindset of individuals that depends on behavior, feelings, attitudes as a difference of each human being's characteristics [1]. Personality often resembles a person's behavior and unique characteristic. The Big Five Personality Traits and Myers Briggs Type Indicator (MBTI) are commonly used for personality analysis [2]. The Big Five Personality Traits are acknowledged as an effective way to identify a person's personality because it is more informative [3]. The Big Five Personality model can predict personality for any benefits, namely for prospective job applicants analysis and relationship matching analysis [4]. Social media has been operated as a platform to share moods, feelings, thoughts, and issues about their situation on daily life. Twitter as a social media has been excessively active and prominent in Indonesia. It is convenient to collect, store and analyze Twitter Users' personalities based on their Tweets [5].

Several studies have been fulfilled to identify social media users' personalities based on texts or tweets in their account. Previous study by [6], used Naïve Bayes for the classification method, implemented LIWC feature, and the term weighting achieved 53.96% accuracy score. The data set size was 211 Twitter users with 474.888 tweets data. The authors stated that the low accuracy score is due to the imbalanced data. Similar study conducted by [7], implemented Decision Tree C4.5 method and TF-RF and TF-CHI$^2$ as the linguistic approach for personality prediction and achieved 65.72% accuracy score on 145 Twitter users with total of tweets data as many as 331.439 tweets data. The accuracy of 65.72% was obtained by combining social features and TF-RF as the linguistic approach. The authors of this research stated that the low accuracy score is because the inequality of the data, therefore the model tends to predict only on the dominant class of the data. Another experiment was conducted to compare different methods such as SVM, BLR, MNB, and CNN on 250 users with 9900 text data [8]. The performance results from that research showed that the optimal accuracy they achieved was equal to 61.6% with CNN method with LIWC as the linguistic approach. Their optimal performance result was because they extracted LIWC features into CNN model. This research stated that the use of language

Accepted by editor: 08-06-2021 | Final Revision: 26-07-2021 | Online Publication: 20-08-2021

features and words features is suitable for understanding personality traits. Personality detection research usually implements Linguistic Inquiry Word Count (LIWC) as a linguistic approach, it is practical to analyze a person's personality based on a text document and it can increase the performance of the system [8].

Yusra *et al.* [9] applied Naïve Bayes method to classify personality traits on 1500 tweets data from 15 users using the Big Five Personality with a satisfying result. The accuracy score they had was 86.66%. However, the dataset size they used was very small. This research did not implement any linguistic approach, but the labelling process was labelled by a psychologist. The data they had was 95 users, due to the inequality of data they only used 15 users. Previous research applied Support Vector Machine (SVM) to identify personality using text classification process with 8660 text data and achieve up to 88.40% accuracy score [10]. This research compared three different methods: SVM, Naïve Bayes, and Neural Networks. Their results stated that SVM method has the best performance comparing to Naïve Bayes and Neural Networks. The SVM method achieved 84.78% accuracy score, Naïve Bayes achieved 75.85% accuracy score, and Neural Networks achieved 69.8%. The optimal accuracy score was also obtained by combining TF-IDF and LIWC as the linguistic approach. According to several research, SVM has a better performance results compared to other methods for personality prediction system.

Another research related to personality detection implemented a text-semantic approach to gain better performance result [11]. The semantic approach that can be implemented is Bidirectional Encoder Representations (BERT), this research gains 92% accuracy score in classification task [12]. According to this research [13], BERT pre-trained model is effective as a semantic approach for the classification task. BERT also has been implemented on sentiment analysis with 82% accuracy score [14]. On previous research, BERT has only been used for text classification such as sentiment analysis, it has not been implemented on personality prediction yet. According to those several research, BERT pre-trained model might be suitable for the semantic approach to perform the personality prediction system. BERT as the state-of-art model implements semantic information of the context in text data [12]. One of BERT pre-trained models namely, “IndoBERT” is a model which effective in semantic approach that achieves great performance for NLP classification tasks [13]. Previous research on sentiment analysis combined Bayesian Network with BERT and it stated that it increases the performance of the system [15]. Therefore, in this research we will experiment on combining BERT with SVM method but in personality prediction system.

This research is combining SVM with BERT as the semantic approach for personality prediction system. The dataset size in this research is larger than previous researches with 295 Twitter users. Our main reference is a previous study by [7], the accuracy scores they obtained are still low. Their research only uses social features data with term weighting as the linguistic approach. Hence, this research will implement LIWC as the linguistic approach because it is more relevant and capable to extract the linguistic feature from text [8]. The authors also stated that the gap in their research was because of the imbalance data [6]. As for the solution to the imbalance data problem, this research will implement SMOTE as an oversampling technique. The previous research by [7] and [6], had only experimented on one method. Therefore, the innovation of this research is we conducted an experiment to combine SVM and SMOTE with BERT as the semantic approach on personality prediction system. According to [11], implementation of semantic approach has the ability to improve the performance of the system. The purpose of this research is to analyze the impact of oversampling technique (SMOTE), linguistic feature (LIWC), and semantic approach (BERT) on personality prediction system on Twitter users using Big Five Personality model.

The rest of the paper is organized as follows. Section 2 describes the research method of personality prediction system on Twitter. Section 3 provides the results and discussion of experiments and followed by the conclusion in Section 4.

## 2. Research Method

The system plan of the personality prediction system is shown in Figure 1, which consists of data crawling, labelling, pre-processing, implementing LIWC features, classification process with, and finally evaluating the performance

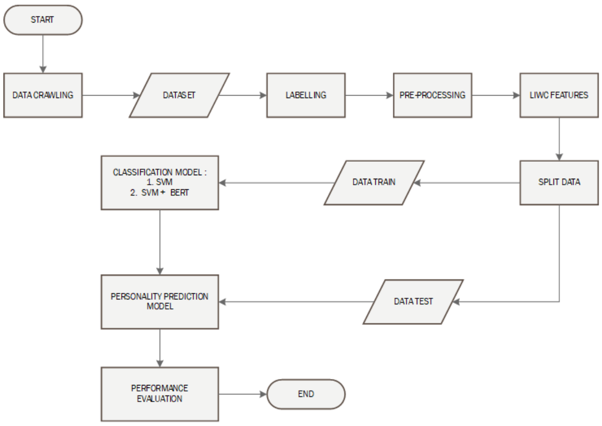

Figure 1. Personality Prediction System

DOI: https://doi.org/10.29207/resti.v5i4.3197

### 2.1. Big Five Personality

Big Five Personality Trait is a personality model with simple implementation to understand and commonly used for predicting personality. Big Five Personality has five dimensions: Openness, Conscientiousness, Extrovert, Agreeableness, and Neuroticism. Every dimension has different characters and meanings [16].

Openness personality trait is a person with high curiosity, active imagination, attentiveness, and caring characteristic. Conscientiousness personality trait is a person with careful, cautious, wide-awake, and thorough character. Extraversion personality trait is a friendly, easy-going, talkative, cheerful, passionate, and enthusiastic person. Agreeableness personality trait have characteristics such as high sympathy, attentive, and most likely to work in a team. Neuroticism personality trait is a person with anxious, nervous, self-doubt and frustrated character. Linguistic aspect tends to have a significant influence on determining personality [8].

### 2.2. Data Crawling

Data crawling is an extraction process to collect data from a website that can be stored and analyzed [6]. Twitter data is collected by a crawling data system which has been developed in previous research [17]. The features that we collected from Twitter users are the social features data as shown in Table 1 which consists of username, sum of following, sum of followers, sum of tweets, sum of URLs, sum of media URLs, sum of retweets, sum of hashtags, sum of mentions, sum of punctuations, and sum of uppercases. The total of data is 511,617 tweets data from 295 Twitter users.

Table 1. Social Features Data Descriptions

| Social Features | Descriptions |
|---|---|
| Sum of Follower | The number of user's followers |
| Sum of Following | The number of user's following |
| Sum of Tweets | The number of user's tweets |
| Sum of URLs | The number of URLs users had shared |
| Sum of Media URLs | The number of media URLs users had shared |
| Sum of Retweets | The number of user's retweets |
| Sum of Hashtags | The number of user's hashtags |
| Sum of Mentions | The number of user's mentions |
| Sum of Punctuations | The number of punctuations users had used |
| Sum of Uppercases | The number of uppercases users had used on Twitter |

### 2.3. Data Labelling

The labelling process will make uses of Big Five Inventory (BFI) questionnaire which has been developed on previous research [18]. This questionnaire is accountable and has been used for determining Big Five Personality on several research [7]. The questionnaire consists of 25 questions with 5 questions for each personality trait. The answer for each question is represented in scale from 1 to 5. Scale 1 represents strongly disagree and Scale 5 represents strongly agree.

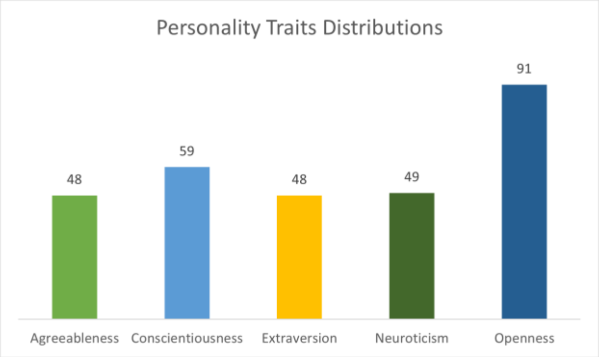

Figure 2. Twitter Users Personality Traits Distribution

The questionnaire has been distributed through Twitter users and this research has collected 295 Twitter users. There are 91 users with openness trait, 59 users with conscientiousness trait, 48 users with extraversion trait, 48 users with agreeableness trait, and 49 users with neuroticism trait. The distribution for each personality trait is shown in Figure 2.

### 2.4. Text Preprocessing

Text preprocessing is one of the steps that need to be accomplished before analyzing data [10]. This step is essential to obtain a better quality of the data. There are six steps of preprocessing which consist of data cleaning, case folding, word tokenization, data normalization, stop words, and stemming.

Data cleaning is a process to remove symbols, URLs, and numbers from a sentence. Case folding is a process to convert all the letters to lowercase in a sentence. Word tokenization is a process to separate words by a space into a token. Data normalization is a process to normalize all the uncommon words with various types of writings to formal words. Stop words is used to remove all the unnecessary words which has no meaning in a sentence. Stemming is the last step for preprocessing to return all the words to a basic form by removing suffix, infix, and prefix. In this text preprocessing we implement “Sastrawi” as the Python library for stop words and stemming process.

### 2.5. Linguistic Inquiry Word Count

Linguistic Inquiry Word Count (LIWC) is a method to count words automatically based on their categories [19]. Pennebaker has developed LIWC since 2007. There are two features of LIWC, namely open vocabulary and closed vocabulary. The closed vocabulary feature can analyze the correlation between language and psychological variables [10]. Table 2 shows the correlation scores between the LIWC category and Big Five Personality that has been developed on previous research [6]. The closed

DOI: https://doi.org/10.29207/resti.v5i4.3197

vocabulary feature is defined by collecting words category based on LIWC, which has a significant correlation value. The vocabulary is gathered from the official LIWC website by translating the vocabulary to a formal Indonesian language [6].

Table 2. LIWC Correlation Scores

| LIWC Category | O | C | E | A | N |
|---|---|---|---|---|---|
| 1st person | -0.19 | 0.02 | 0.03 | 0.08 | 0.10 |
| 2nd person | -0.16 | 0 | 0.16 | 0.08 | -0.15 |
| 3rd person | -0.06 | -0.08 | 0.04 | 0.08 | 0.02 |
| 1st person plural | -0.10 | 0.03 | 0.11 | 0.18 | -0.07 |
| Pronouns | -0.21 | -0.02 | 0.06 | 0.11 | 0.06 |
| Negations | -0.13 | -0.17 | -0.05 | -0.03 | 0.11 |
| Assent | -0.11 | -0.09 | 0.07 | 0.02 | 0.05 |
| Prepositions | 0.17 | 0.06 | -0.04 | 0.07 | -0.04 |
| Numbers | 0.08 | 0.04 | -0.12 | 0.11 | -0.07 |
| Affect | -0.12 | -0.06 | 0.09 | 0.06 | -0.12 |
| Positive Emotion | -0.11 | -0.02 | 0.11 | 0.14 | 0.01 |
| Negative Emotion | 0 | -0.18 | 0.04 | -0.15 | 0.16 |
| Anxiety | -0.2 | -0.05 | -0.03 | -0.03 | 0.17 |
| Anger | 0.3 | -0.19 | 0.03 | -0.23 | 0.13 |
| Sadness | -0.3 | -0.11 | 0.02 | 0.01 | 0.10 |
| Discrepancy | -0.12 | -0.13 | -0.07 | -0.04 | 0.13 |
| Tentative | -0.06 | -0.10 | -0.11 | -0.07 | -0.12 |
| Certainty | -0.06 | -0.10 | 0.10 | 0.05 | 0.13 |
| Seeing | -0.04 | -0.01 | -0.03 | 0.09 | -0.01 |
| Hearing | -0.08 | -0.12 | 0.12 | 0.01 | 0.02 |
| Feeling | -0.01 | -0.05 | 0.06 | 0.10 | 0.10 |
| Communication | -0.06 | -0.07 | 0.13 | 0.02 | 0 |
| Friends | -0.01 | 0.06 | 0.15 | 0.11 | -0.08 |
| Family | -0.17 | 0.05 | 0.09 | 0.19 | -0.07 |
| Humans | -0.09 | -0.12 | 0.13 | 0.07 | -0.05 |
| Time | -0.22 | 0.09 | 0.02 | -0.12 | 0.01 |
| School | 0.02 | 0.04 | -0.07 | -0.01 | 0.06 |
| Job/work | 0.04 | 0.07 | -0.08 | -0.07 | 0.07 |
| Achievement | -0.05 | 0.14 | -0.09 | 0.05 | 0.01 |
| Home | -0.20 | 0.50 | 0.03 | 0.19 | 0 |
| Sports | -0.14 | 0 | 0.05 | 0.06 | -0.01 |
| Tv/movies | 0.05 | 0.06 | 0.05 | -0.05 | -0.02 |
| Music | 0.04 | -0.11 | 0.13 | 0.08 | -0.02 |
| Money/finance | -0.04 | -0.08 | -0.04 | -0.11 | 0.04 |
| Metaphysical | 0.07 | -0.08 | 0.08 | -0.01 | -0.01 |
| Death | 0.15 | -0.12 | 0.01 | -0.13 | 0.03 |
| Religion | 0.05 | -0.04 | 0.11 | 0.06 | -0.03 |
| Sexuality | 0 | -0.06 | 0.17 | 0.08 | 0.03 |
| Eating/drinking | -0.15 | -0.04 | 0.18 | 0.03 | -0.01 |
| Sleep | -0.14 | -0.03 | 0.02 | 0.11 | 0.10 |
| Grooming | -0.20 | -0.05 | -0.01 | 0.07 | 0.05 |
| Swear words | 0.06 | -0.14 | 0.06 | -0.21 | 0.11 |

## 2.6. Bidirectional Encoder from Transformers Representations (BERT)

BERT is an effective semantic approach and a pre-trained language model which works based on a bidirectional transformer [15]. This experiment implements a BERT pre-trained model, namely "IndoBERT". IndoBERT model is specifically made for the dataset in Indonesian language [13]. BERT has been very successful on several Natural Language Process (NLP) tasks [20].

BERT makes use of Transformer to learn contextual relation among words in a text. The Transformer itself consists of an encoder and a decoder. Encoder reads text input and decoder writes or produces the prediction for classification tasks. On BERT, the encoder reads the whole sequence of words on the text input at once. It is considered bidirectional because it reads the sequence both ways, left-to-right and right-to-left. The illustration of how BERT works is shown in Figure 3.

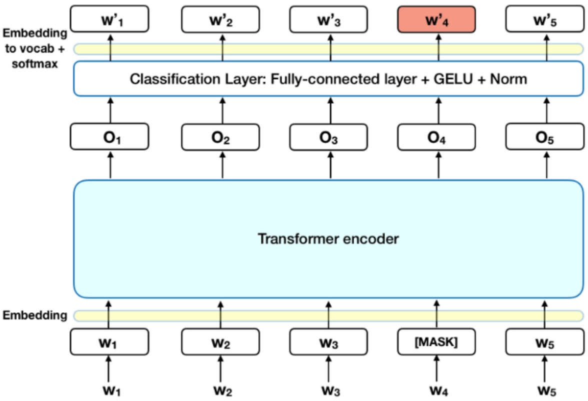
w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Figure 3. Illustration of BERT [21]

There are three tokens [CLS], [SEP], and [MASK]. CLS token is placed at the beginning of the sentence and SEP token is placed at the end of the sentence. There are 15% of words in a sequence that will be replaced with MASK token. The model will figure out the original word of the masked word based on the other non-masked word by adding a classification layer on the top of encoder output, transforming output vectors into a vocabulary dimension, and finally using SoftMax for calculating the probability of each word in the vocabulary.

## 2.7. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a linear model for classification and regression that can solve linear and non-linear problem [22]. SVM model separates the data into classes by creating a line called a hyperplane.

SVM algorithm finds the best hyperplane by finding the closest distance between points and the line from the classes. These points are called the support vectors and the distance is called the margin. This classification method aims to maximize the margin to find the optimal hyperplane as shown in Figure 4.

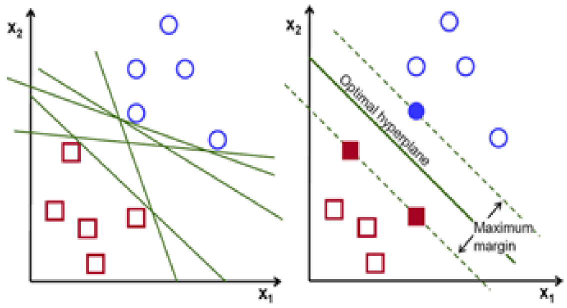

Figure 4. Illustration of Optimal Hyperplane on SVM [23]

As for the optimization, on the classification process this research implements Synthetic Minority Oversampling Technique (SMOTE). The SMOTE optimization is beneficial for handling imbalanced data and allows increasing the quality of the SVM Classifier [22]. There are several variations of SMOTE namely, SMOTE, SMOTE-NC, Borderline-SMOTE, and SVM-SMOTE.

In this research we implement SVM-SMOTE. Based on previous research, this method is more effective for the data imbalance problem compared to other oversampling methods because it only focuses on the borderline area due to the fact that this area is the most critical area for establishing decision boundary [24]. The SVM-SMOTE works by approximating the borderline area by the support vectors after the training process on SVM Classifier and synthetic data will be created along the lines joining each minority class randomly.

2.8. Support Vector Machine combined with BERT

The main idea of the classifier based on SVM and BERT is combining the SVM method and BERT as the semantic approach. As we can see in Figure 3, on BERT Classifier we implement embedding first before fitting it into the classification process.

In SVM and BERT Classifier, we also implement that embedding process. As for the classification layer, we replace the classifier with the SVM classifier. The flowchart of this method is shown in Figure 5. In this method, we use the same data train and data test. This method also implements "IndoBERT" pre-trained model. As for parameters in BERT model is using Batch Size = 32, Learning Rate = 3e-5, and number of epochs = 4. These parameters are based on previous research which stated that these selected parameters are the optimal values on various classification tasks [25].

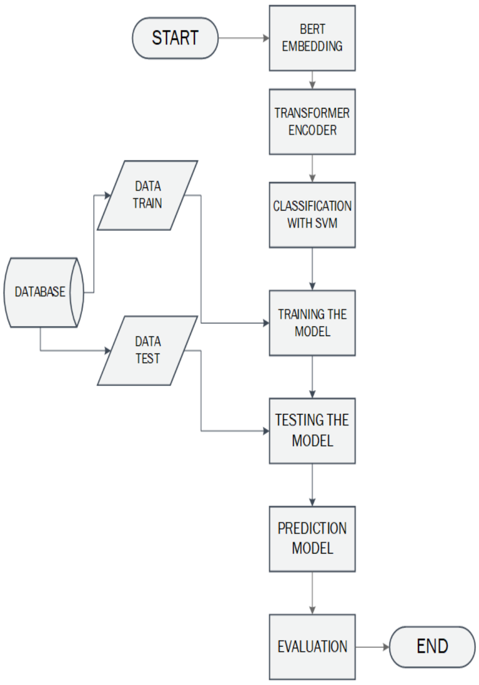

Figure 5. SVM and BERT Prediction System

## 3. Result and Discussion

This research was carried out by three scenarios. The first scenario used SVM method as the baseline and compared it with the combination of SVM with SMOTE method. The comparison was conducted to optimize the model and used for handling the imbalance data. The second scenario determined the best feature to use for the model, therefore we compared social features with social features combined with linguistic features (LIWC). The last scenario compared SVM plus SMOTE with SVM plus SMOTE and BERT using the best features that we have already determined on the previous scenario. The total of the data set size is 295 Twitter users with 511,617 tweets data and it is divided into 80% of data train and 20% of data test. The data was labelled by using BFI questionnaire to categorize each Twitter account into the Big Five personality trait.

### 3.1. Results

The result of the first scenario that compares between SVM method and SVM with SMOTE is shown in Table 3. The result has shown that the implementation of SMOTE with SVM has the better accuracy score than using SVM method stand alone. The result of using SVM only gain 57.97% accuracy score and after the implementation of SMOTE with SVM gained higher accuracy score as equal to 77.65%. This has shown that SMOTE is effective as the optimization for handling the imbalanced data. The accuracy score after the implementation of SMOTE is better compared to the one without SMOTE.

Table 3. First Scenario Results using SVM and SMOTE

| Personality Traits | Accuracy | |
|---|---|---|
| | SVM | SVM + SMOTE |
| Agreeableness | 83.05% | 79.41% |
| Conscientiousness | 42.37% | 76.47% |
| Extraversion | 25.42% | 80.88% |
| Neuroticism | 81.36% | 75.00% |
| Openness | 57.63% | 76.47% |
| Average Accuracy | 57.97% | 77.65% |

As stated by the previous scenario, SVM with SMOTE has the better performance, hence in the second scenario we used SVM with SMOTE. The second scenario was conducted to find out the best feature to implement in this model. This scenario compared between social feature and the combination of social feature with linguistic feature (LIWC). The result of the second scenario is shown in Table 4.

Table 4. Second Scenario Result using LIWC Feature

| Personality Traits | Accuracy | |
|---|---|---|
| | Social Feature | LIWC + Social Feature |
| Agreeableness | 79.41% | 76.46% |
| Conscientiousness | 76.47% | 79.41% |
| Extraversion | 80.88% | 88.23% |
| Neuroticism | 75.00% | 77.94% |
| Openness | 76.47% | 75.01% |
| Average Accuracy | 77.65% | 79.41% |

The implementation of SVM with SMOTE using social feature achieved 77.65%. On the other hand, using the implementation of linguistic feature (LIWC) combined with social feature achieved 79.41% accuracy score. There is an increase as equal to 1.76%. This has shown that the addition of LIWC feature has a significant impact as the linguistic aspects for the model. This stated that combination of LIWC feature with social feature is better than only using social feature for personality prediction system.

According to previous research, the best feature to implement is the combination of linguistic feature (LIWC) and social feature. Therefore, in the third scenario we implement the combination of linguistic feature (LIWC) and social feature. The third scenario compared between SVM with SMOTE method and the addition of BERT as the semantic approach as shown in Table 5.

Table 5. Third Scenario using SVM and SMOTE with BERT

| Personality Trait | Accuracy | |
|---|---|---|
| | SVM + SMOTE | SVM + SMOTE + BERT |
| Agreeableness | 76.46% | 80.88% |
| Conscientiousness | 79.41% | 79.41% |
| Extraversion | 88.23% | 86.76% |
| Neuroticism | 77.94% | 77.94% |
| Openness | 75.01% | 75.36% |
| Average Accuracy | 79.41% | 80.07% |

The results of the third scenario shown that the addition of semantic approach (BERT) has the better accuracy than only using SVM and SMOTE. The performance result of SVM and SMOTE with BERT achieved accuracy score as equal to 80.07%. Meanwhile, the accuracy of SVM and SMOTE only achieved 79.41% accuracy score. There is an increase between these two methods as equal to 0.66%.

### 3.2. Discussion

A previous research by [6], used Naïve Bayes method with the implementation of LIWC feature and TF-IDF as the term weighting method achieved 53.96% accuracy score. The accuracy score of this research still has low performance. It may occur because of the imbalanced data they had for classification.

The imbalance data creates bias, it makes the classification model tends to predict only the majority class. As we can see in Figure 2, due to the limitation in this research, the data in this research is also imbalanced. However, the improvement conducted in this research by using oversampling technique namely SMOTE. The implementation of SMOTE has the ability to resample the data especially for the minority data, in this case the minority data is the “Extraversion” class. The “Extraversion” class had the smallest accuracy score based on the result shown in Table 3, the accuracy score for class “Extraversion” only gained 25.42%. After the addition of SMOTE as the optimization for handling imbalanced data, the accuracy score of class “Extraversion” increases up to 80.88%. The implementation of SMOTE makes the data more balance and prevent bias in the classification model. This proved that SMOTE is capable to handle the imbalance data problem and gain a better performance for the system.

In other research by [7], implemented Decision Tree C4.5 method achieved 65.72% accuracy score. The size of the data they used was 145 Twitter users with approximately 300,000 tweets data. As for the linguistic features they used was unigram approach and then calculating it with term weighting namely, TF-RF and TF-CHI$^2$. The accuracy score of the previous research still low may occur because unigram approach did not precisely represent as a linguistic feature for personality prediction.

Linguistic Inquiry Word Count (LIWC) as linguistic feature is capable to count correlation score between words in sentences based on Big Five categories. As shown in Table 4, the addition of LIWC feature in this research could increase the accuracy score of the system. This is because LIWC as linguistic feature is capable to recognize personality traits on users from tweets data, by analyzing the correlation between tweets and psychological variables. The LIWC features has its correlation scores which play quite big role in

DOI: https://doi.org/10.29207/resti.v5i4.3197

determining users' personality based on words they used in their tweets. Based on the analysis we can confirm that the addition of LIWC features have significant impact for achieving a better performance.

Table 6. Comparison of Experiment's Results

| Conditions | Accuracy |
|---|---|
| SVM (Baseline) | 57.97% |
| Baseline + SMOTE | 77.65% (+0.34) |
| Baseline + SMOTE + LIWC | 79.41% (+0.37) |
| Baseline + SMOTE + BERT + LIWC | 80.07% (+0.38) |

Based on these experiment's results on Table 5, the implementation of BERT as the semantic approach combined with SVM method and SMOTE could increase the accuracy score. The improvement of the accuracy score happened because of the embedding process on BERT as the semantic approach. It is also because the BERT model we used is a pre-trained model namely IndoBERT. It is stated on previous research by [13], that this pre-trained model has been trained in approximately 23GB of text data in Indonesian language. Hence, the model already learned multiple times and familiar with the tweets provided because it is in Indonesian language. As shown in Table 6, the conditions that significantly affect the accuracy score is baseline plus SMOTE, BERT, and LIWC with the highest increase towards baseline.

## 4. Conclusion

This research has proposed to implement the combination of SVM with SMOTE, LIWC, and BERT as the semantic approach to predict personality on Twitter users. The performance results of the system are good for dataset as many as 295 Twitter users with 511,617 tweets data. The implementation of semantic approach is the key to improve the performance of the system.

The implementation of SVM with SMOTE, LIWC, and BERT as the semantic approach has shown better performance results comparing to only using SVM method. The semantic approach has shown significant impact in performance results because the BERT model has been trained before and it is more applicable to understand the words in the sentences. Hence, it is concluded that the implementation of BERT as the semantic approach positively affect the personality prediction system to achieved better performance. Applying personality prediction that classifies Twitter users into Big Five Personality traits can make it easier for recruiters to analyze their potential employee's personality through social media.

The limitation of this research is the dataset size still small, this research only gathered 295 Twitter users due to the difficulties of collecting respondents to fill out the BFI questionnaire. The larger amount of data has the possibility to achieve a better performance result. Therefore, future research can improve the performance of the personality prediction system by collecting more respondents for the data and experiment on different methods such as combining BERT with deep learning method for the better performance of the personality prediction system.

## References

[1] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, "Personality Detection from Text using Convolutional Neural Network," *1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019*, vol. 2019, no. Icasert, pp. 1–6, 2019, doi: 10.1109/ICASERT.2019.8934548.

[2] R. Moraes, L. L. Pinto, M. Pilankar, and P. Rane, "Personality Assessment Using Social Media for Hiring Candidates," *2020 3rd Int. Conf. Commun. Syst. Comput. IT Appl. CSCITA 2020 - Proc.*, pp. 192–197, 2020, doi: 10.1109/CSCITA47329.2020.9137818.

[3] F. Celli and B. Lepri, "Is big five better than MBTI? A personality computing challenge using Twitter data," *CEUR Workshop Proc.*, vol. 2253, 2018.

[4] M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, and A. Shakya, "Personality Traits Analysis from Facebook Data," *ICSEC 2017 - 21st Int. Comput. Sci. Eng. Conf. 2017, Proceeding*, vol. 6, pp. 153–156, 2018, doi: 10.1109/ICSEC.2017.8443932.

[5] C. Li, J. Wan, and B. Wang, "Personality Prediction of Social Network Users," *Proc. - 2017 16th Int. Symp. Distrib. Comput. Appl. to Business, Eng. Sci. DCABES 2017*, vol. 2018-Septe, pp. 84–87, 2017, doi: 10.1109/DCABES.2017.25.

[6] F. Ilzam Nur Haq and E. Budi, "Implementasi Naive Bayes Classifier untuk Prediksi Kepribadian Big Five pada Twitter Menggunakan Term Frequency-Inverse Document Frequency ( TF-IDF ) dan Term Frequency-Relevance Frequency ( TF-RF ) Program Studi Sarjana Ilmu Komputasi Fakultas Informatik," *e-Proceeding Eng.*, vol. 6, no. 2, pp. 9785–9795, 2019.

[7] Willy, E. B. Setiawan, and F. N. Nugraha, "Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter," *2019 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Big Data Artif. Intell. IC3INA 2019*, pp. 114–119, 2019, doi: 10.1109/IC3INA48034.2019.8949601.

[8] C. Yuan, J. Wu, H. Li, and L. Wang, "Personality Recognition Based on User Generated Content," *2018 15th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2018*, pp. 1–6, 2018, doi: 10.1109/ICSSSM.2018.8465006.

[9] Yusra *et al.*, "Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naïve Bayes," no. November, pp. 2579–5406, 2018.

[10] S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 1076–1082, 2018, doi: 10.1109/ICACCI.2018.8554828.

[11] M. Hassanein, "Predicting Personality Traits from Social Media using Text Semantics," *2018 13th Int. Conf. Comput. Eng. Syst.*, pp. 184–189, 2018.

[12] W. Li, S. Gao, H. Zhou, Z. Huang, K. Zhang, and W. Li, "The automatic text classification method based on bert and feature union," *Proc. Int. Conf. Parallel Distrib. Syst. - ICPADS*, vol. 2019-Decem, pp. 774–777, 2019, doi: 10.1109/ICPADS47876.2019.00114.

[13] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020, [Online]. Available: http://arxiv.org/abs/2009.05387.

[14] M. G. Sousa, K. Sakiyama, L. D. S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, "BERT for stock market sentiment analysis," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2019-Novem, pp. 1597–1601, 2019, doi:

DOI: https://doi.org/10.29207/resti.v5i4.3197

10.1109/ICTAI.2019.00231.

[15] S. Liu, H. Tao, and S. Feng, “Text Classification Research Based on Bert Model and Bayesian Network,” *Proc. - 2019 Chinese Autom. Congr. CAC 2019*, pp. 5842–5846, 2019, doi: 10.1109/CAC48633.2019.8996183.

[16] Y. J. Nie, G. J. Gao, Y. X. Wang, D. X. Liu, and K. Gao, “Personality predicting model based on user’s linguistic behavior,” *Proc. 2017 9th Int. Conf. Model. Identif. Control. ICMIC 2017*, vol. 2018-March, no. Icmic, pp. 827–832, 2018, doi: 10.1109/ICMIC.2017.8321569.

[17] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, “Data Crawling Otomatis pada Twitter,” no. September, pp. 11–16, 2016, doi: 10.21108/indosc.2016.111.

[18] R. R. Mccrae *et al.*, “The NEO – PI – 3 : A More Readable Revised NEO Personality Inventory The NEO – PI – 3 : A More Readable Revised NEO Personality Inventory,” *J. Pers. Assess.*, vol. 84, no. 3, pp. 261–270, 2016, doi: 10.1207/s15327752jpa8403.

[19] İ. Ergu, “Twitter Verisi ve Makine Ö ğ renmesi Modelleriyle Ki ş ilik Tahminleme Predicting Personality with Twitter Data and Machine Learning Models,” no. 1, 2019.

[20] Z. Gao, A. Feng, X. Song, and X. Wu, “Target-dependent sentiment classification with BERT,” *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[21] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, “A survey of word embeddings for clinical text,” *J. Biomed. Informatics X*, vol. 4, no. April, p. 100057, 2019, doi: 10.1016/j.yjbinx.2019.100057.

[22] L. Demidova and I. Klyueva, “SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem,” *2017 6th Mediterr. Conf. Embed. Comput. MECO 2017 - Incl. ECYPS 2017, Proc.*, no. June, pp. 17–20, 2017, doi: 10.1109/MECO.2017.7977136.

[23] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. G. Nieto, A. B. Sánchez, and M. M. Fernández, “Hard-rock stability analysis for span design in entry-type excavations with learning classifiers,” *Materials (Basel).*, vol. 9, no. 7, pp. 1–19, 2016, doi: 10.3390/ma9070531.

[24] A. C. Flores, K. D. Gorro, R. I. Icoy, and C. F. Peña, “An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set,” *2018 Int. Conf. Eng. Appl. Sci. Technol.*, pp. 1–4, 2018.

[25] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

DOI: https://doi.org/10.29207/resti.v5i4.3197