Received: November 4, 2020. Revised: December 23, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model

Veronica Ong[1] Anneke D. S. Rahmanto[1] Williem Williem[1]
Nicholaus H. Jeremy[1] Derwin Suhartono[1]* Esther W. Andangsari[2]

[1]*Computer Science Department, School of Computer Science,*
Bina Nusantara University, Jakarta, 11480, Indonesia
[2]*Psychology Department, Faculty of Humanities, Bina Nusantara University, Jakarta 11480, Indonesia*
* Corresponding author's Email: dsuhartono@binus.edu

Abstract: Over the past few years, there has been an increasing number of researches on automated personality prediction. One of the approaches include analysing personality based on the user's choice of words on social media. Even though research on personality prediction have been done in several different languages, advancements in personality prediction for Bahasa Indonesia, the Indonesian language, has been stagnant. This is due to scarcity of data and the lack of psycholinguistic dictionaries for the language. This is unfortunate as Indonesians are among one of the most active social media consumers, considering more than 2% of worldwide tweets come from Indonesians. We address these issues in this study through the modelling of our own personality prediction model based on an Indonesian dataset that we have produced. We employed the model using the XGBoost machine learning algorithm, trained on 250 user data that we have collected and annotated manually. The resulting model was able to gain decent performance for the Agreeableness and Openness personality trait, achieving AUROC of 0.71 and 0.63, while the Conscientiousness, Extraversion, and Neuroticism personality traits were harder to distinguish with an AUROC of 0.5, 0.59, and 0.48 respectively.

Keywords: Bahasa Indonesia, Five factor model, Personality datasets, Personality prediction, Twitter.

1. Introduction

The growth of the Internet has promoted the ease of accessibility to social media worldwide. In fact, [1] reported that one in every four minutes of time on the Internet is spent on social media. Another research showed that users login to Facebook 2 to 5 times a day, with each session ranging approximately five to fifteen minutes [2]. Intense usage of social media can also be seen among Indonesians. [3] reported that 8.23% of the world's tweets come from Jakarta (capital city of Indonesia) users. Additionally, Mr. Roy Simangunsong, Twitter Country Head of Indonesia, mentioned that the daily percentage of active Indonesian users on Twitter is 77% [4].

Social media as a huge data source, provides an opportunity for researchers to perform data mining. Relevant studies have been introduced, including sentiment analysis [5-7], named entity recognition [8-10], automatic summarization [11-13], and user profile clustering [14].

Another area of research with social media as a data source is personality prediction, which is the focus of this study. The intuition behind this research is the way social media functions as a platform for users to share information. In fact, one of the seven-block framework on the functions of social media is the identity block, which represents how users consciously or unconsciously reveal their identities on social media [15]. Although much research on personality prediction has been done in the past, the task is still rarely explored thoroughly for the Indonesian language. This study aims to establish the methods employed to perform personality prediction in Bahasa Indonesia – from dataset creation to user personality modelling.

The contributions of this study are as follows.

- The present work presents a new Indonesian Twitter dataset that may be used for personality prediction. The dataset contains data from 250 users, each labelled with the Five Factor Model personality traits based on annotations by psychology experts.
- Personality prediction models based on XGBoost for each personality trait – Agreeableness, Extraversion, Conscientious-ness, and Neuroticism.

This study is an extension and improvement from the previous research [16], with the following advancements:

- The current study uses a dataset where each user was inter-rated by three psychology experts, whereas the previous study only involved one psychology expert per user. This was done to ensure objectivity of annotation and decrease individual bias of an expert.
- Additional feature engineering was added to the current study, such as tweeting likelihood, tweeting steadiness, etc.

This study consists of the following: Section 2 discusses the relevant work that has been done on personality prediction. Section 3 describes the adopted methodology for the current study. Section 4 presents the result findings and analysis of this study. Section 6 consists of the conclusion.

2. Literature review

2.1 The five factor model

The Five Factor Model is a hierarchical structure of personality traits which consists of 5 main dimensions: Agreeableness, Extraversion, Conscientiousness, Neuroticism (its opposite is often referred to as Emotional Stability), and Openness to Experience [17].

Agreeableness assesses an individual based on his/her quality of interpersonal perspective [18]. Extraversion evaluates the quantity, depth and need of one's interpersonal interaction. Conscientiousness looks at a person's level of organization, persistence, and motivation towards a certain goal. Neuroticism assesses the emotional stability of an individual in response to psychological distress. Openness to Experience scores an individual based on his/her susceptibility and tolerance to new or unfamiliar experiences.

Previous studies have shown correlation between social media use and an individual's Five Factor personality. As reported in [19], highly extraverted individuals tend to belong to more Facebook groups compared to introverted individuals. This study also reported that users with high neurotic trait tend to have the Wall as their favorite Facebook component, whereas those low in neuroticism favored photos. Users with higher Openness to Experience have a greater tendency to be sociable in Facebook. Additionally, a similar study was conducted by [20] with several other findings. The study reveals a tendency in higher number of friends amongst highly extroverted users than users in the less extroverted group. Those with high extroversion also exhibited lower tendency to share personal information. Unlike the previous finding by [19], this study reported that neurotic users showed a tendency to post photos on their profile. Users with high agreeableness exhibited behavior of using less page features compared to users with low agreeableness. Highly conscientious users displayed a tendency of higher number of friends and showed less usage of picture upload feature.

2.2 Previous works on social media personality prediction

[21] conducted the study using Twitter data, relying on Twitter user behaviour features as well as psychological dimension features through predefined dictionaries such as Linguistic Inquiry and Word Count (LIWC) and MRC Psycholinguistic Database. A similar study by [22] employed publicly available user Twitter behaviours and influence scores as features to their model [23] took a different approach by fully taking advantage of n-gram features instead of predefined dictionaries to perform personality prediction on bloggers. Similarly, [24] made use of n-grams from myPersonality dataset. Their research takes a step further by adding topic features, representing subject matters that users tend to post about on Facebook. The study by [25] similarly employed textual features for their model on a Twitter dataset.

In contrast to the aforementioned studies implemented using supervised learning, [26] proposed an unsupervised approach exploiting language-independent features based on LIWC and MRC. A semi-supervised learning approach by [27] was attempted based on Twitter data, while [28] attempted it based on Facebook data. Recent advancements in deep learning has also led to development of deep learning-based personality assessment models as attempted by [29].

Apart from English, similar related work has also been applied to other languages. An experiment for personality prediction in Chinese by [30] trained a model using Sina Weibo social network data with

Table 1. List of information extracted from twitter

Twitter user information	**User information**
1. Number of tweets The total number of tweets posted by a user on the time of data extraction. 2. Number of followers The total number of Twitter accounts following a user on the time of data extraction. 3. Number of following The total number of Twitter accounts followed by a user on the time of data extraction. 4. Number of favorites The total number of times a user favorited a certain tweet. 5. Number of retweets amongst extracted tweets The number of times a user posted a retweet amongst the extracted tweets 6. Number of retweeted tweets amongst extracted tweets The number of times a user's tweet was retweeted amongst the extracted tweets. 7. Number of quotes amongst extracted tweets The number of times a user posted a quote tweet amongst the extracted tweets. 8. Number of mentions amongst extracted tweets The number of times a user used mentions amongst the extracted tweets. 9. Number of replies amongst extracted tweets The number of times a user posted a reply tweet amongst the extracted tweets. 10. Number of hashtags amongst extracted tweets The number of times a user used hashtags amongst the extracted tweets. 11. Number of URLs amongst extracted tweets The number of times a user posted a link on a tweet amongst the extracted tweets.	1. Date and time of tweet The date and time when the tweet was posted. 2. Tweet text The tweet content posted by a user. 3. Type of tweet The type of tweet posted by a user (normal, retweet, quote, reply) • Normal tweet: type of tweet where user merely posted a status without tagging other users or reposting content from other users. • Retweet: type of tweet where a user reposted a tweet by another Twitter user, with no additional comments. • Quote: type of retweet where a user added additional comments regarding what was being retweeted. • Mention tweet: type of tweet where some user tagged other Twitter users in a tweet without referencing any previous tweets. • Reply tweet: type of tweet where some user tagged other Twitter users in a tweet while referencing a certain existing tweet. 4. Content of tweet user replied to The content of the tweet that is being replied by a user, if the type of tweet is a reply tweet.

user behavior and LIWC features. [31] also attempted their study on Sina Weibo data, however using other lexicons (Tongyici Cilin, Hownet and ITH). Research in Chinese was extended by [32] based on Facebook data from Chinese users through the use of non-psycholingusitic features such as n-gram features and user behaviour. Besides Chinese, a relevant study has also been applied to Russian language by [33] based on their native social media, VKontakte. Additionally, [34] focused on classifying Twitter user personality in the Indonesian language, by using translated version of the myPersonality dataset.

Based on these studies, we observed that advancements in personality prediction for languages other than English are stagnant, particularly for Indonesian language. Two major problems were identified on why this was the case:

- Gold standard dataset for personality prediction is in English.
- Most studies rely on psycholinguistic features, such as LIWC and MRC, which are readily available for several languages, such as English and Chinese, but are not available for Indonesian language.

To address these problems, in this study, we present a new Indonesian dataset for personality prediction with a machine learning model that leverages non-psycholinguistic features as an alternative.

3. Methodology

The workflow of the study is divided into 6 parts, mainly data collection, data preparation, data annotation, feature engineering, and training of personality classifiers. The overall workflow of this study is provided in Fig. 1.

Table 2. List of information extracted from twitter

No	Elements removed from the data	Purpose/consideration of element removal
1	Omitting and counting retweets	A retweet isn't an original post by the Twitter user, thus it doesn't reflect the language use of a Twitter user. However, the act of retweet itself might reveal behavioral tendencies of a personality trait. Hence, the number of retweets amongst the extracted tweets are retained.
2	Replace mention with "[UNAME]" token	Twitter mentions comply to a certain format: "@" as the first character, followed by 15 or less alphanumeric or underscore characters (e.g. @username). Mentions reflect the digital identity of Twitter users and are used in tweets to tag other Twitter users. Since mentions are commonly used but is unique to each Twitter user, it is replaced with the "[UNAME]" token.
3	Replace hashtag with "[HASHTAG]" token	Similar to mentions, hashtags on Twitter also use a certain format: "#" is used to represent the first character of a hashtag, followed by any character except spaces and punctuation (e.g. #topic). While hashtags can be an indicator for a user's topic of preference, its free-form nature can result in many unique tokens. For this reason, it is replaced with the "[HASHTAG]" token.
4	Remove hyperlinks	Any kind of link shared on Twitter are automatically converted into a certain format: "https://t.co/code", where "code" comprises of 10 randomly generated alphanumeric characters. As hyperlinks follows a fixed format, these are removed from the tweet.
5	Remove emojis	The program built to extract tweets currently doesn't support emojis as input.
6	Remove non-Bahasa Indonesia tweets	Non-Bahasa Indonesia tweets are removed to maintain the scoping of this study.
7	Remove tweets containing only "[UNAME]" or "[HASHTAG]"	Tweets that only contain "[UNAME]" or "[HASHTAG]" doesn't represent much significance or context.
8	Remove tweets with elements from other social media	Elements from third-party social media (e.g. Path) are removed as some of the elements are generated by the third parties, and thus doesn't reflect the user's language use on Twitter.
9	Remove empty tweets	Removal of the previously mentioned elements can sometimes result in an empty tweet. Such occurrences are removed from the dataset.
10	Remove excessive number of tweets	Due to limited resources, the authors and psychology experts decided to provide a maximum of 100 tweets per Twitter user.
11	Remove users with less than 20 tweets	The experts expressed that 20 tweets or less would be too little to assess the user's personality. Hence, Twitter users with only 20 tweets or less are dropped from the dataset.

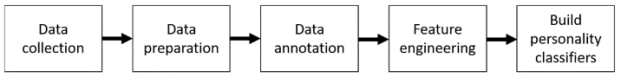

Figure. 1 Workflow of the study

3.1 Data collection

Twitter data was gathered from October 2016 to December 2016 using the Twitter API. User tweets and Twitter user information were extracted from each user. The extracted data covers several information as listed on Table 1.

Users were chosen based on two main criteria. Firstly, the user shows active behaviour of Twitter by posting at least once per month. This criterion helps to ensure the user's behavioural consistency on Twitter (i.e. avoiding drastic behaviour change due to user's long inactivity on Twitter). Secondly, the user tweets in Bahasa Indonesia as their main language. Not all Indonesian Twitter users adopt Bahasa Indonesia as their main tweeting language, so this second criterion is important to maintain the scoping of the current study. For each user, a maximum of 200 tweets were extracted.

In this study, a total of 250 data were collected, where each data represents one Twitter user. The current study aims to improve the annotation method compared to the previous study [16], which is elaborated on section 3C. However, due to limited resources, the current study only leveraged a subset of 250 data from the original 359 data.

3.2 Data preparation

Following extraction, a short set of data element removal was done for each user's tweets content. These element removals were chosen in agreement between the authors and the psychology experts who agreed to participate in the study. Table 2 presents the list of the data element removal functions that were applied in this study, as well as the purpose or consideration for each function. The dataset that has gone through element removal functions from Table 2 are passed to each psychology expert for annotation.

3.3 Data annotation

Each user's collected data were passed to three psychology experts. The psychology experts that have been chosen to participate in this study are those who are familiar with The Five Factor Model, have at least a master's degree in psychology, and are faculty members of the academic institution's psychology department.

Every user is given a "high" or "low" scoring for each Five Factor Model personality trait, namely Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. Indicators that determine whether a user has high or low level of a certain personality trait is based on the Handbook of Personality Assessment by Weiner and Greene [35]. Thus, the data annotation process resulted in 15 labels per user – 5 scores ("high" or "low") from each of the three psychology experts who participated in this study.

Finally, the resulting annotations from three psychology experts were consolidated into a single and final annotation through majority vote. This final annotation serves as the predicted variable in the dataset for training and evaluating the machine learning algorithm. The distribution of the combined dataset is shown on Table 3.

Table 3. Final annotation distribution

Personality	High	Low
Agreeableness	117	133
Conscientiousness	31	219
Extraversion	190	60
Neuroticism	63	187
Openness	115	135

3.4 Feature engineering

In addition to the features mentioned in Table 1, several other features were added by aggregating data from user tweets.

- **Average time difference.** This feature measures the average time difference between each tweet. The time difference is calculated by subtracting the posting datetime of a certain tweet, and its preceding tweet. The average time difference is the average value from the time differences calculated for each pair of tweets.
- **Tweeting likelihood of the day.** This measures the fraction of extracted tweets that were posted on weekdays or weekends. This feature was similarly used in [36] where the authors predict response likelihood of information solicitation from social media data.
- **Tweeting steadiness.** This feature measures the steadiness of a user's tweeting frequency. Tweeting steadiness is defined as σ, where σ is the standard deviation of the time difference between each tweet. This feature was also used in [36].
- **Fraction of tweets that are quote.** This feature measures the proportion of quote tweets amongst the extracted tweets. This is calculated by dividing the number of quote tweets by the total number of extracted tweets.
- **Fraction of tweets that are replies or mentions.** This feature measures the proportion of reply or mention tweets amongst the extracted tweets. The proportion of reply or mention tweets are calculated by dividing the number of reply and mention tweets by the total number of extracted tweets.
- **Average tweet length.** This feature represents the average length of a tweet in characters.
- **Word diversity (number of unique unigrams).** This feature measures the lexical diversity of a user, by calculating the number of unique unigrams/tokens used amongst the extracted tweets.
- **Preference of replied topic.** This feature represents the preferred topic that the user tends to reply to, based on their reply content (if it is a reply tweet). Topics are generated using the LDA (Latent Dirichlet Allocation) topic modelling algorithm with k = 10 using Gibbs sampling. Each tweet that the user replied to serves as the document used in the LDA algorithm. LDA results are further elaborated on section 4B.

The result of the feature engineering step is a dataset appended with the mentioned additional features for each Twitter user.

Furthermore, several functions were implemented to normalize the use of slang words. Normalization functions are performed to reduce the number of unique unigrams. Normalization functions were adapted from [37, 38] with slight modification. To preserve the user's preferences in using slang words, additional variables are added for every normalization function to count the number of times

Table 4. Number_convert conversion rules

Numbers	Conversion
0	o
1	i
2	Copy char/string before "2"
3	e
4	a
5	s
6	g
7	t
8	b
9	g

Table 5. Fix_spelling conversion rules

Abnormal form	Normal form
oe	u
dj	j

Table 6. Fix_spelling conversion rules

Abnormal form	Normal form
-ny	-ny
-nk	-ng
-x	-nya
-z	-s
-dh	-t

a function is performed. The 6 normalization functions applied in this study are as follows.

- check_kbbi
  Checks whether a given term exists in KBBI, the official dictionary of the Indonesian language. This function precedes all other normalization functions. The number of terms that exist in KBBI are stored as an additional feature named exists_in_kbbi.
- number_convert
  Replace the occurrence of number(s) to proper letters. Conversion rules are presented in Table 4. The number of times number_convert is performed on a given user's tweets are stored as the need_number_convert feature.
- fix_spelling
  Fix abnormal terms which are slightly modified to imitate spelling of Indonesia words similar to that of the old Indonesian language. The number of times fix_spelling is performed on a given user's tweets are stored as the need_fix_spelling feature. The fix_spelling conversion rules are shown on Table 5.
- fix_suffix
  Fix abnormal terms which are slightly modified to imitate spelling of Indonesia words similar to that of the old Indonesian language. The number of times fix_spelling is performed on a given user's tweets are stored as the need_fix_spelling feature. The fix_spelling conversion rules are shown on Table 6.
- remove_repeat
  Trim abnormal terms with at least 3 adjacent repeating letters (e.g. "looooh" to "loh"). The number of times remove_repeat is performed on a given user's tweets are stored as the need_remove_repeat feature.
- exists_in_alaydict
  This function replaces abnormal words to their normal form based on a list of words stored in Alay Dictionary (slang word dictionary). This dictionary is generated from [38]'s research. The dictionary is a mapping between an abnormal term and its normal form. The number of terms that exist in Alay Dictionary are stored as the exists_in_alaydict feature.

3.5 Build personality classifiers

This step involves building the personality prediction classifier. The machine learning algorithm is trained using the resulting dataset that has been preprocessed, annotated by psychology experts, and appended with additional engineered features. The current study utilized gradient boosted trees called XGBoost, due to the promising results delivered in the previous study using the same algorithm [16].
Five personality binary classifiers were built for this use case – one for each Five Factor Model personality trait. The resulting dataset from the previous step was split into two. The first split serves as the training dataset to train the machine learning model, while the second split as the testing dataset to evaluate the trained machine learning model. Class distributions of each split are presented on Table 7 and Table 8.

Table 7. Class distribution for first split (train dataset)

Personality	High	Low
Agreeableness	107	93
Conscientiousness	172	28
Extraversion	49	151
Neuroticism	150	51
Openness	102	98

Table 8. Class distribution for second split (test dataset)

Personality	High	Low
Agreeableness	26	24
Conscientiousness	47	3
Extraversion	11	39
Neuroticism	37	12
Openness	33	17

Table 9. Hyperparameter tuning list

Hyperparameter form	Hyperparameter value
Subsample ratio of training instances	0.5, 0.75, 1
Subsample ratio of columns for each split	0.6, 0.8, 1
Number of trees	1000
Learning rate	0.01, 0.001, 0.0001
Gamma	1
Minimum sum of instance weight in a child node	1

Model training was performed using 3-fold cross validation and a set of tuning parameters, as listed on Table 9.

4. Results and discussion

4.1 Behaviorial analysis

An analysis was done to observe relevancy between personality traits and Twitter user behavior. The Spearman correlation's ρ estimation was calculated to inspect the monotony of this relationship. Top 10 correlated behaviors for each personality trait are presented in Table 10. Correlation coefficient values ≥ 0.3 and ≤ -0.3 are bolded.

Some moderate correlations can be observed between the Agreeableness trait and URL count. Conscientious users are found to be correlated with longer tweet length and fewer terms that require need_remove_repeat normalization. The dataset also shows some relevancy between the Extraversion personality trait and a user's number of reply tweets.

4.2 Word analysis

A separate analysis was done to explore and understand the characteristics of the dataset's tweet contents. There are two types of tweet contents that were considered for the analysis – the contents of tweets posted by the user and the contents of tweets which the user replied to.

Two types of word analysis were done in this study, namely topic modelling and bag of words.

1) Topic modelling

The contents of tweets which the user replied to were compiled into a separate dataset to explore latent topics. This reveals the general topics that a Twitter user tends to engage with through replying. Topics were generated using the Latent Dirichlet Allocation algorithm with k = 10 using Gibbs sampling. A visualization of the result is presented in Fig. 2 by picking terms with the highest β values.

By choosing k = 10, several closely related words can be observed from several topics.

For instance, topic 1 demonstrates the use of modernized slang words and slang particles used at the end of a sentence such as:

- "selfie" (self photo)
- "*elu*" (you)
- "*neh*" (particle to emphasize an object)
- "*yha*" (particle to emphasize expression)
- "*lh*" (particle to emphasize expression)

This pattern can also be seen in topic 2, with the occurrence of informal words such as:

- "*gue*" (me)
- "*lo*" (you)

Slang particles can also be spotted in this topic. Topic 9 also displays a similar trend with slang words such as:

- "mention" (mention, a feature in Twitter that allows the tagging of other users)
- "*neng*" (informal greeting towards women)
- "*ngopi*" (drinking coffee)

Another attribute that these topics have in common are the occurrences of emoticons.

Topic 5 reveals more emotional terms compared to other topics, such as

- "*terima kasih*" (thank you)
- "*happy*" (happy)
- "*selamat*" (congratulations)
- "*kangen*" (missing you)
- "*semoga*" (hoping/wishing)
- "*sukses*" (good luck)

These terms can usually be used to engage in conversations with other people. Another emotional term here is *"banget"*, which is a slang word usually used to exaggerate one's emotions (e.g. *"happy banget"* would translate to very happy).

Political tendencies can be observed from topic 3, due to words like:

- "dpr" (DPR, the official legislative body of Indonesia)
- "*gubernur*" (governor)
- "*presiden*" (president)

A similar trend is found on topic 8. Examples of political terms are:

- "ahok" (name of Jakarta's former governor)
- "*pemimpin*" (leader)
- "anies" (name of Jakarta's current governor)
- "*politik*" (politics)
- "*partai*" (political party)

A subtler political trend can also be found in topic 6 through words like:

Table 10. Spearman correlation between each personality trait and twitter user behavior

Personality	Behavior	Spearman Correlation
Agreeableness	**URL count**	**0.3217**
	Average of time between each tweet	0.2850
	Hashtag count	0.2599
	Followers count	0.2589
	Mentions count	0.2452
	Standard deviation of time between each tweet	0.2424
	Tweet steadiness	-0.2424
	Quote count	-0.1927
	Proportion of quotes	-0.1913
	Proportion of tweets posted on Saturday	0.1877
Conscientiousness	**Average tweet length**	**0.3398**
	Number of times need_remove_repeat normalization function is performed	**-0.3099**
	Reply count	-0.2459
	Number of times tweet term exists in KBBI	-0.2318
	Mention and reply count	-0.2292
	Proportion of replies	-0.2292
	Average of time between each tweet	0.1965
	Tweet count	-0.1852
	Number of tweet term that exists in Alay Dictionary	-0.1812
	Standard deviation of time between each tweet	0.1805
Extraversion	**Reply count**	**0.3470**
	Mention and reply count	0.2991
	Proportion of replies	0.2991
	Tweet count	0.2820
	Average of time between each tweet	-0.2654
	Tweet steadiness	0.2585
	Standard deviation of time between each tweet	-0.2585
	Average tweet length	-0.2231
	Followers count	0.1934
	Favorites count	0.1866
Neuroticism	Followers count	-0.1593
	Number of times tweet terms exist in KBBI	-0.1385
	Proportion of unique unigrams	-0.1288
	Retweets count	0.1275
	Hashtag count	-0.1231
	Number of tweets posted on weekday	0.1207
	Tweet count	-0.1157
	Mentions count	-0.1083
	URL count	-0.1034
	Number of tweets posted on Saturday	0.0977
Openness	Retweeted count	0.2791
	Number of tweets posted on Saturday	0.1808
	Mention and reply count	-0.1746
	Proportion of replies	-0.1746
	Number of unigrams	0.1737
	Proportion of tweets posted on Saturday	0.1663
	Number of unique unigrams	0.1631
	Replies count	-0.1557

- “jokowi" (name of the seventh and current president of Indonesia)
- “mulyani” (name of Indonesia’s current Minister of Finance)

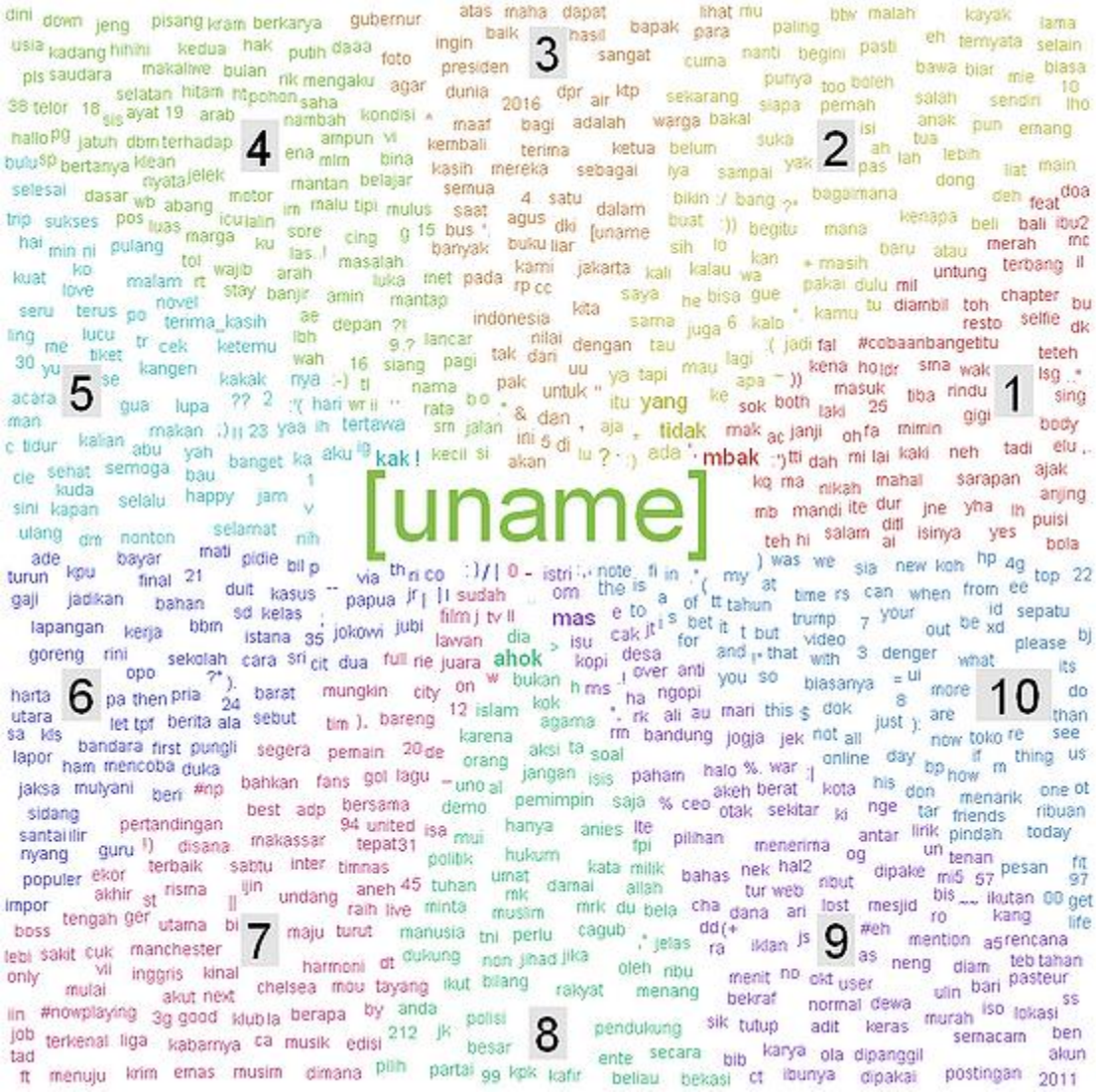

Figure. 2 Visualization of topic modelling result

- "kpu" (KPU, the general elections commission of Indonesia)

Topic 7 exhibits athletic terms, especially those related to soccer. Examples of these terms include:

- "timnas" (Indonesia national football team)
- "*gol*" (goal)
- "*pemain*" (pemain)
- "*liga*" (league)
- "*pertandingan*" (competition)

Words like *"manchester"*, *"united"*, *"chelsea"* suggests the name of several famous professional football clubs.

Finally, topic 10 stands out as the cluster with the most English words.

After generating topics through LDA, we inspected the replying tendencies of each personality trait (i.e. what kinds of topics does a certain personality trait tend to reply to). Table 11 shows the correlations between each topic and personality trait, where correlation coefficients ≥ 0.3 and ≤ -0.3 are bolded. Results suggests that the highly agreeable users tend to avoid reacting to topics 6 and 8, which are both about politics. Moderate relevancy was also shown between highly extraverted users and reactive response towards topic 5.

2) Bag of words

Each user's tweets were tokenized into unigrams to be used as features. Relevancy between each unigram and personality trait were calculated as shown in Table 12. Correlation coefficient values ≥ 0.3$ and ≤ -0.3 are bolded.

Results show that highly agreeable users tend to avoid mentioning the words:

- "ahok" (name of Jakarta's former governor)
- "*katanya*" (he/she said)
- "*bang*" (informal honorific for young man)

Moderate correlation was also observed in which highly conscientious people are related to the higher

Table 11. Spearman correlation between each personality trait and LDA topic

Personality trait	**Topic No.**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Agreeableness	-0.2069	-0.2517	-0.2847	-0.2238	-0.0471	**0.3341**	-0.1544	**0.4548**	-0.2994	-0.1441
Conscientious-ness	-0.2411	-0.2576	-0.1130	-0.1990	-0.2825	-0.1292	-0.2190	-0.1347	-0.2140	-0.1956
Extraversion	0.2135	0.2751	0.0733	0.1895	0.3166	-0.0008	0.1783	-0.0051	0.2007	0.1990
Neuroticism	-0.0047	0.0174	0.0566	0.0862	-0.0434	0.0392	-0.0346	0.1345	-0.0046	0.0498
Openness	-0.1074	-0.1150	-0.1462	-0.1140	-0.0871	-0.1742	-0.0489	-0.2190	-0.1558	-0.0458

Table 12. Spearman correlation between each personality trait and LDA topic

Personality	Term	Spearman correlation	Personality	Term	Spearman correlation
Agreeableness	**ahok**	**-0.4381**	Conscientiousness	**menjadi**	**0.3574**
	hari	**0.3566**		**untuk**	**0.3553**
	katanya	**-0.3442**		**dominan**	**0.3389**
	bang	**-0.3105**		**kebangkitan**	**0.3359**
	mui	-0.2910		**merayakan**	**0.3328**
	agama	-0.2836		**menggunakan**	**0.3302**
	with	0.2764		**telah**	**0.3301**
	terima_kasih	0.2690		**solusi**	**0.3197**
	isu	-0.2638		**diperlukan**	**0.3189**
	demo	-0.2585		**membuka**	**0.3174**
Extraversion	**bangsa**	**-0.3493**	Neuroticism	agama	0.2506
	swt	-0.2861		ba	0.2461
	[uname]	0.2843		islam	0.2415
	kekayaan	-0.2790		pki	0.2366
	meraih	-0.2684		warga	0.2352
	jika	-0.2673		agamanya	0.2352
	pintar	-0.2564		kelakuan	0.2336
	pemilu	-0.2486		marah	0.2285
	kepada	-0.2481		saatnya	0.2200
	persoalan	-0.2459		dihina	0.2197
Openness	**selalu**	**0.3019**			
	dan	0.2757			
	day	0.2609			
	di	0.2426			
	semoga	0.2417			
	baik	0.2398			
	sukses	0.2369			
	hari	0.2364			
	ahok	-0.2360			
	tak	0.2347			

Table 13. Training AUROC descriptive statistics

Personality trait	AUROC minimum	AUROC mean	AUROC maximum	AUROC median	AUROC std dev
Agreeableness	0.6502774	0.7704263	0.8400324	0.7752603	0.04250346
Conscientiousness	0.6148551	0.7992447	0.8774977	0.8183706	0.05565995
Extraversion	0.5773775	0.7482653	0.8058905	0.7628676	0.05148066
Neuroticism	0.4945098	0.6011416	0.6745098	0.5996078	0.03932261
Openness	0.6077967	0.7046526	0.7584577	0.7153149	0.03355072

use of relatively wise words such as:

- "*dominan*" (dominant)
- "*kebangkitan*" (revival)
- "*merayakan*" (celebrate)
- "*solusi*" (solution)
- "*diperlukan*" (need)

Highly extraverted users demonstrate a lesser tendency with political terms. However, they show a subtle higher use of "[uname]", a term converted from Twitter's "@", which is used to tag other Twitter users. Neuroticism doesn't display strong associations with terms, but top correlated terms are mostly related to politics and religion. Finally, users with high openness demonstrated a higher usage of well intention terms, such as:

- "*selalu*" (always)
- "*semoga*" (hopefully)
- "*baik*" (good)
- "*sukses*" (good luck or success)

4.3 Prediction model evaluation

Each personality trait classifier was trained using preprocessed and feature engineered data through 135 combinations of set hyperparameters. Training evaluation is performed using 3-fold cross validation using the AUROC (Area Under the ROC Curve) metric. Table 13 shows the descriptive statistics of the training AUROC for each classifier, while the resulting optimized tuning parameters for each personality trait classifier are presented in Table 14.

Table 14. Optimized tuning parameters for each personality trait classifiers

Personality	Hyperparameter		Personality	Hyperparameter	
Agreeableness	Number of trees	: 1000	Conscientiousness	Number of trees	: 1000
	Maximum tree depth	: 4		Maximum tree depth	: 2
	Learning rate	: 0.01		Learning rate	: 0.01
	Gamma	: 1		Gamma	: 1
	Subsample ratio of columns for each split	: 1		Subsample ratio of columns for each split	: 0.6
	Minimum sum of instance weight in a child node	: 1		Minimum sum of instance weight in a child node	: 1
	Subsample ratio of training instances	: 0.5		Subsample ratio of training instances	: 0.5
Extraversion	Number of trees	: 1000	Neuroticism	Number of trees	: 1000
	Maximum tree depth	: 4		Maximum tree depth	: 8
	Learning rate	: 0.01		Learning rate	: 0.01
	Gamma	: 1		Gamma	: 1
	Subsample ratio of columns for each split	: 0.6		Subsample ratio of columns for each split	: 0.6
	Minimum sum of instance weight in a child node	: 1		Minimum sum of instance weight in a child node	: 1
	Subsample ratio of training instances	: 0.5		Subsample ratio of training instances	: 0.5
Openness	Number of trees	: 1			
	Maximum tree depth	: 8			
	Learning rate	: 0.01			
	Gamma	: 1			
	Subsample ratio of columns for each split	: 1			
	Minimum sum of instance weight in a child node	: 1			
	Subsample ratio of training instances	: 0.5			

Optimized models were then used to predict a held-out dataset. Table 15 shows the AUROC of each personality classifier. Results show that the agreeableness classifier performed the best with an AUROC of 0.713, whereas the conscientiousness and neuroticism classifier performed poorly with AUROC below and equal to 0.5.

Poor performance on conscientiousness classifier could probably be due to very imbalanced distribution between classes. When inspected, the classifier predicted all test dataset instances as "low" class. The AUROC difference between training and training also suggests that the conscientiousness model is overfitted. Meanwhile, the reason behind the neuroticism classifier's poor performance could be due to lack of more significant predictors, as there were no strong associations between features and predictor as shown in Table 10, Table 11, and Table 12. As seen in Table 13, the neuroticism classifier also consistently displayed the worst performance across all descriptive statistics compared to other classifiers.

5. Conclusions

In this study, we have attempted personality prediction in the Indonesian language using a collection of 250 Twitter users' data. The personality prediction models were trained using a new Indonesian labelled dataset that was presented in this study. The XGBoost models were able to reach decent results, with AUROC of 0.71 for the Agreeableness trait, and 0.63 for the Openness trait. Several moderate associations were found between each personality traits and user behaviour on Twitter and their choice of language for all traits except Neuroticism. Moreover, the Neuroticism personality prediction model performed the worst, which suggests that the dataset for Neuroticism still needs to be improved upon.

By inspecting users' replies to other tweets, we managed to find several distinguishable topics such as politics, slang words, emotional terms, and sports. These topics served as a user's reaction to different kinds of engagements on Twitter.

In the future, this research plans to build up the dataset in terms of size and reliability. Weak associations and poor performance for several personality traits suggest that there is still a large potential in building the dataset.

Table 15. AUROC of personality classifiers

Personality Trait	AUROC
Agreeableness	0.7131410
Conscientiousness	0.5000000
Extraversion	0.5909091
Neuroticism	0.4864865
Openness	0.6274510

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, D.S. and V.O.; methodology, D.S., V.O., A.D.S.R., and W.; software, V.O.; validation, A.D.S.R., and W.; formal analysis, D.S.; investigation, V.O.; resources, E.W.A.; data curation, E.W.A.; writing—original draft preparation, V.O. and N.H.J.; writing—review and editing, N.H.J.; visualization, V.O.; A.D.S.R.; supervision, D.S.; project administration, D.S.; funding acquisition, D.S.

Acknowledgments

The authors would like to thank Dr. Aryo E. Nugroho and Dr. Muhamad N. Suprayogi for their participation in labelling the personality traits of the dataset used in this study, along with one of the authors in this study, Dr. Esther W. Andangsari.

This research was funded by Ministry of Research, Technology and Higher Education of the Republic of Indonesia under "Penelitian Produk Terapan" grant number 039A/VR.RTT/VI/2017.

References

[1] J. Mander, GWI Social 2016 - Summary Report. Available: http://www.slideshare.net/globalwebindex/globalwebindex-social-q1-summary-report, 2015

[2] A. Quan-Hasse and A. L. Young, "Uses and Gratifications of Social Media: A Comparison of Facebook and Instant Messaging", *Bulletin of Science, Technology & Society*, Vol. 30, No. 5, pp. 350-361, 2010.

[3] StatCounter GlobalStats. *Social Media Stats Indonesia*. Available: https://gs.statcounter.com/social-media-stats/all/indonesia, 2020.

[4] A. Maulana, *Twitter Rahasiakan Jumlah Pengguna di Indonesia*. Available: https://www.cnnindonesia.com/teknologi/20160322085045-185-118939/twitter-rahasiakan-jumlah-pengguna-di-indonesia, 2016

[5] A. Farzindar and D. Inkpen, "Natural Language Processing for Social Media", *Synthesis Lectures on Human Language Technologies*, pp. 1-166, 2015.

[6] P. Melville, V. Sindhwani, and R. D. Lawrence, "Social Media Analytics: Channeling the Power

of the Blogosphere for Marketing Insight", In: *Proc. of the WIN*, Vol. 1, No. 1, pp. 1-5, 2009.

[7] S. Yuliyanti, T. Djatna, and H. Sukoco, "Sentiment Mining of Community Development Program Evaluation Based on Social Media", *TELKOMNIKA (Telecommunication, Computing, Electronics, and Control)*, Vol. 15, No. 4, pp. 1858-1864, 2017.

[8] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B. S. Lee, "Twiner: Named Entity Recognition in Targeted Twitter Stream", In: *Proc. of 35th International ACM SIGIR Conf. on Research and Development in Information Retrieval*, Portland, OR, pp. 12-16, 2012.

[9] A. Purwarianti, L. Maldberger, and M. Ibrahim, "Supervised Entity Tagger for Indonesian Labor Strike Tweets Using Oversampling Technique and Low Resource Features", *TELKOMNIKA (Telecommunication, Computing, Electronics, and Control)*, Vol. 14, No. 4, pp. 1462-1471, 2016.

[10] A. Ritter, S. Clark, and O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study", In: *Proc. of Conf. on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 2008.

[11] W. Hariardi, N. Latief, D. Febryanto, and D. Suhartono, "Automatic Summarization from Indonesian Hashtag on Twitter Using TF-IDF and Phrase Reinforcement Algorithm", In: *6th International Computer Science and Engineering*, WCSE, Tokyo, Japan, pp. 17-19, 2016.

[12] M. Kågebäck, O. Morgen, N. Tahmasebi, and D. Dubhashi, "Extractive Summarization Using Continuous Vector Space Models", In: *2nd Workshop on CVSC*, pp. 31-39, 2014.

[13] H. Lin and J. Bilmes, "A Class of Submodular Functions for Document Summarization", In: *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1*, Portland, OR, 2011.

[14] C. Virmani, A. Pillai, and D. Juneja, "Clustering in Aggregrated User Profiles Accross Multiple Social Networks", *International Journal of Electrical and Computer Engineering*, Vol. 7, No. 6, pp. 1-16, 2017.

[15] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social Media? Get Serious! Understanding Functional Building Blocks of Social Media", *Business Horizons*, Vol. 54, No. 3, pp. 241-251, 2011.

[16] V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi, "Personality Prediction Based on Twitter Information in Bahasa Indonesia", In: *Proc. of Federated Conf. on Computer Science and Information Systems*, Prague, Czech Republic, 2017.

[17] R. R. McCrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications", *Journal of Personality*, Vol. 60, No. 2, pp. 175-215, 1992.

[18] D. Cervone and L. A. Pervin. *Personality: Theory and Research*, 12th ed. Hokoben, NJ: John Wiley & Sons, 2013.

[19] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and Motivations Associated with Facebook Use", *Computers in Human Behavior*, Vol. 25, No. 2, pp. 578-586, 2009.

[20] Y. Amichai-Hamburger and G. Vinitzky, "Social Network Use and Personality," *Computers in Human Behavior*, Vol. 26, pp. 1289-1295, 2010.

[21] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting Personality from Twitter" In: *PASSAT and 2011 IEEE Third Conference on SocialCom*, Boston, MA, pp. 9-11, 2011.

[22] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter", In: *Proc. of the 2011 IEEE Third International Conf. on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conf. on Social Computing*, Boston, MA, 2011.

[23] F. Icaobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large Scale Personality Classification of Bloggers", In: *ACII 2011*, pp. 568-577, 2011.

[24] H. A. Schwarz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawak, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach", *PloS One*, Vol. 8, pp. 3692-3699, 2013.

[25] A.V. Kunte and S. Panicker, "Using textual data for personality prediction: a machine learning approach", In: *Proc. of 2019 4th International Conf. on Information Systems and Computer Networks (ISCON)* pp. 529-533, IEEE. 2019.

[26] M. Poesio, PR2: A Language Independent Unsupervised Tool for Personality Recognition from Text. [Online]. Available: arXiv:1402.2796 [cs.CL], 2014.

[27] A. C. E. S. Lima and L. N. De Castro, "A Multi-Label Semi-Supervised Classification Approach Applied to Personality Prediction in Social

Media", *Neural Networks*, Vol. 58, pp. 122-130, 2014.
[28] H. Zheng and C. Wu, "Predicting Personality Using Facebook Status Based on Semi-supervised Learning", In *Proc. of the 11th International Conf. on Machine Learning and Computing*, pp. 59-64, 2019.
[29] J. Zhao, D. Zeng, Y. Xiao, L. Che, and M. Wang, "User personality prediction based on topic preference and sentiment analysis using LSTM model", *Pattern Recognition Letters*, 138, pp. 397-402, 2020.
[30] D. Wan, C. Zhang, M. Wu, and Z. An, "Personality Prediction Based on All Characters of User Social Media Information", In: *Proc. of the Chinese National Conf. on Social Media Processing*, Beijing, China, 2014.
[31] S. Han, H. Huang, and Y. Tang, "Knowledge of words: An interpretable approach for personality recognition from social media", *Knowledge-Based Systems*, 105550, 2020.
[32] K. H. Peng, L. H. Liou, C. S. Chang, and D.S. Lee, "Predicting Personality Traits of Chinese User Based on Facebook Wall Posts", In: *Proc. of 24th Wireless and Optical Communication Conf.*, Taipei, Taiwan, 2015.
[33] M. Stankevich, A. Latyshev, N. Kiselnikova, and I. Smirnov, "Predicting Personality Traits from Social Network Profiles", In: *Proc. of Russian Conf. on Artificial Intelligence* (pp. 177-188). Springer, Cham, 2019.
[34] B. Y. Pratama and R. Sarno, "Personality Classification Based on Twitter Text Using Naïve Bayes, KNN, and SVM", In: *Proc. of International Conf. on Data and Software Engineering*, Yogyakarta, 2015.
[35] I. B. Weiner and R. L. Greene, *Handbook of Personality Assessment*, Hoboken, NJ: John Wiley & Sons, 2017.
[36] J. Mahmud, M. X. Zhou, N. Megiddo, J. Nichols, and C. Drews, "Recommending Targeted Strangers from Whom to Solicit Information on Social Media", In: *Proc. of the International Conf. on Intelligent User Interfaces*, Santa Monica, CA, 2013.
[37] G. A. Buntoro, T. B. Adji, and A. E. Purnamasari, "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation", In: *Proc. of Conf. on Information Technology and Electrical Engineering*, ed. Hanifah Rahmi Fajrin, pp. 39-43, 2014.
[38] A. R. Naradhipa and A. Purwarianti, "Sentiment Classification for Indonesian Message in Social Media", In: *Proc. of International Conf. on Cloud Computing and Social Networking*, Bandung, West Java, pp. 26-27, 2012.