# Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier

Nur Haziqah Zainal Abidin[1], Muhammad Akmal Remli[2], Noorlin Mohd Ali[3]
Danakorn Nincarean Eh Phon[4], Nooraini Yusoff[5], Hasyiya Karimah Adli[6], Abdelsalam H Busalim[7]
Faculty of Computing, Universiti Malaysia Pahang, 26600, Pekan, Pahang, Malaysia[1, 3, 4]
Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, City Campus
Pengkalan Chepa 16100, Kota Bharu, Kelantan, Malaysia[2, 5, 6]
The Irish Institute of Digital Business, DCU Business School
Dublin City University, Dublin, Ireland[7]

***Abstract*—The term "personality" can be defined as the mixture of features and qualities that built an individual's distinctive characters, including thinking, feeling and behaviour. Nowadays, it is hard to select the right employees due to the vast pool of candidates. Traditionally, a company will arrange interview sessions with prospective candidates to know their personalities. However, this procedure sometimes demands extra time because the total number of interviewers is lesser than the total number of job seekers. Since technology has evolved rapidly, personality computing has become a popular research field that provides personalisation to users. Currently, researchers have utilised social media data for auto-predicting personality. However, it is complex to mine the social media data as they are noisy, come in various formats and lengths. This paper proposes a machine learning technique using Random Forest classifier to automatically predict people's personality based on Myers–Briggs Type Indicator® (MBTI). Researchers compared the performance of the proposed method in this study with other popular machine learning algorithms. Experimental evaluation demonstrates that Random Forest classifier performs better than the different three machine learning algorithms in terms of accuracy, thus capable in assisting employers in identifying personality types for selecting suitable candidates.**

***Keywords*—Machine learning; random forest; Myers–Briggs Type Indicator® (MBTI); personality prediction; random forest classifier; social media; Twitter user**

## I. INTRODUCTION

Machine learning is a well-known technique that is broadly utilised by researchers for personality prediction. Due to the advantages of machine learning in learning historical data and making a prediction on future data, the researcher can also use it for learning personality patterns [1]. Such an application is also well-known in psychological science as an assessment tool to predict personality. Nowadays, businesses and recruiters are investing in personality prediction technologies that utilise the machine learning technique. By developing a machine learning algorithm, selecting the best candidates can be achieved, and an error occurred due to the manual analysis process can be reduced.

Motivational influences and human behaviour are the best predictors in personality that will predict an individual's work performance. People's experiences which are emotionally significant with situations, can also be influenced by personality. This approach reflects a person's character and can identify using the Myers-Briggs Type Indicator (MBTI). Based on [2], they defined the personality of a person as a set of attributes that describes a likelihood on the uniqueness of behaviour, feeling and thoughts of the person. These attributes of a person change through time and positions. In a simpler term, we can regard personality as a mixture of characteristics and standards that built an individual's unique character. There are many different personality models used to characterise personality such as the Big Five model (Five-factor model) [3], Myers-Briggs Type Indicator (MBTI) [4], and Theory of Personality Types Carl Jung [5]. Among these personality models, the Big Five and MBTI models are currently popular among researchers. Compared to other models, MBTI is more robust as it has broader applications in different disciplines, although it suffers some issues in terms of reliability and validity. In this study, we select the MBTI personality model due to its popularity and potential to be utilised in different fields.

People nowadays deliver their thoughts and emotions through social media platforms [6]. The posts can be in so many ways, such as using an image, URL link, and music. People's personality also can be examined using social media. The personality of people shown to be useful in predicting job satisfaction, professional and romantic relationship success. In the process of selecting the right candidates, companies nowadays tend to examine the candidates' social media profiles to know the personality of the candidates for a particular job [7]. They intend to reduce the time spent in the preliminary phases of recruitment which is typically known as social media mining. In this paper, we used Twitter as it is one of the most popular social media platforms used nowadays.

It is not easy for employers to select the best candidates for their companies [8][9]. Furthermore, the traditional procedure usually requires employers to spent time conducting interviews with all shortlisted candidates. With the rapid

development of the internet, some researchers have been developing personality prediction system based on the candidates' social media postings to identify candidates' personality for employers [7] accurately.

Over the past few years, many studies use various machine learning algorithm for predicting personality types. One of the earliest studies on Predicting Personality System from Facebook users was developed in 2017 by Tandera [6]. The goal of this study was to build a prediction system that can automatically predict the users' personality based on their activities on Facebook [6]. They also analysed the accuracy of traditional machine learning and deep learning algorithm on predicting personality by implementing Big Five Personality models. Also, in 2018, Giulio Carducci conducted a study on Computing Personality Traits from Tweets Using Word Embedding and Supervised Learning [10]. [10]. The researcher used a supervised learning approach to compute personality traits from an individual's historical tweets. They developed three machine learning algorithms, namely Support Vector Machine (SVM), LASSO and Logistic Regression to predict Big Five Personality model. Mohammad Hossein Amirhosseini conducted a study on Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator on 2020 [11]. The study developed a new machine learning method for automating the process of meta programme detection and personality type prediction based on the MBTI personality type indicator [11]. The natural language processing toolkit (NLTK) and XGBoost, which is based on Gradient Boosting library in Python is used for implementing machine learning algorithms.

In this paper, an intelligent personality prediction system is proposed to predict the personality of candidates based on Twitter data. The proposed method uses machine learning to mine user characteristics and learn patterns from large amounts of personal behavioural data. This system can automatically evaluate candidates' personality traits by processing various attributes and eliminate time-consuming process required in the conventional approach. The rest of the paper is organised as follows: Section 2 presents related work; Section 3 provides material and methods; Section 4 highlights the result of the experiments, and Section 5 concludes the paper.

## II. Related Works

Many different machine learning algorithms have used by researchers in the study of personality prediction. Almost all research in this field involved several stages, including data gathering, pre-processing, extracting features and perform classification to determine the accuracy of the model. This section highlights the Myers–Briggs Type Indicator® (MBTI) and related works from previous researchers using machine learning algorithms.

### A. Myers–Briggs Type Indicator® (MBTI)

Study on personality has always been a topic of interest for psychologists and sociologists, and one such experiment was performed by the psychiatrist Carl Jung on "Myers-Briggs type indicator". According to [14] in the 1920s, Isabel Myers-Briggs and Katherine Briggs designed the Myers-Briggs type indicator test based on Carl Jung's psychological types. There are 16 personality types on the Myers-Briggs Type Indicator® instrument, which is called a "type table" as shown in Fig. 1 [15]. As an example, someone labelled as INTP in the MBTI system prefers introversion, intuition, thinking and perceiving personality. Based on the label, we can classify the person's desire or behaviour, and more knowledge can be learned by the machine.

The 16 personality types are combined to indicate the personality preferences in four dimensions. Each dimension represents two personalities. The four dimensions are Extroversion–Introversion (E–I), Sensation–Intuition (S–N), Thinking–Feeling (T–F), and Judgment–Perception (J–P) as shown in Fig. 2 [16].

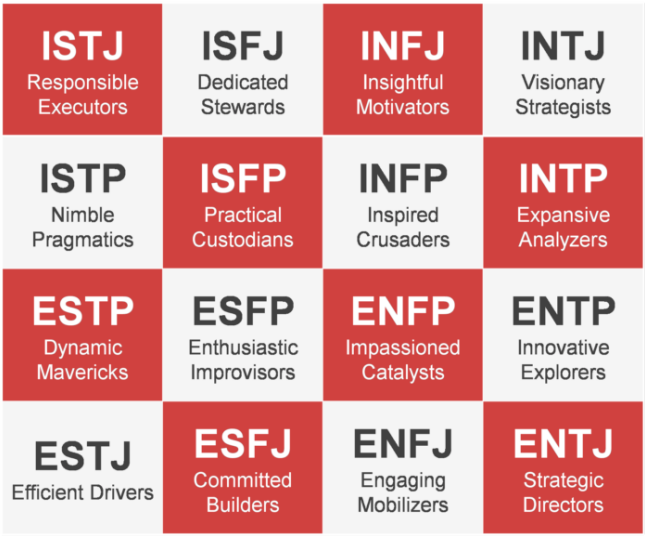


Fig. 1. The Myers Briggs Type Indicator (MBTI).

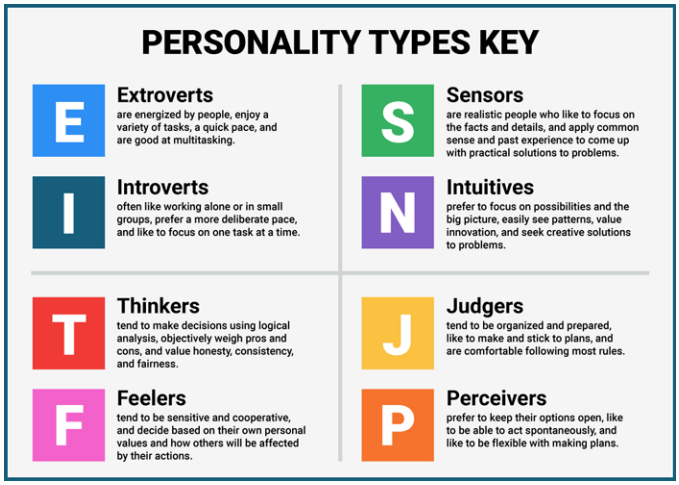


Fig. 2. Key Personality Types.

There are several features related to the various personality types that can be extracted from text or related data. We can use user's posts in social media such as video, image, or other links to analyse their MBTI types using Term Frequency-Inverse Document Frequency (TF-IDF). We can use TF-IDF as a tool to detect and measure the most popular words posted by a person. Beside URL, other potential features can be extracted from text data, including hashtags, emoticons, number of words, ellipses, action words and many more. These extra features also have significant characteristics that could relate to the various personality types. For example, when users of social media are categorised under one of the MBTI, their linguistic contents such as number of words or ellipses will generate extra personality features for the person.

### *B. Machine Learning (ML)*

Machine learning (ML) is a subset of artificial intelligence (AI) that gives frameworks the capacity to naturally take in and improve for a fact without being unequivocally modified [12]. There are three machine learning algorithms which are supervised learning, unsupervised learning and reinforcement learning. The most popular and generally embraced methods are supervised learning and unsupervised learning. Supervised learning is an algorithm that consists of input data (also called training data) and target (or outcome) variable. The input contains a set of features that determine the desired output for the prediction model [13]. Some examples of supervised learning algorithms are Decision Tree, Linear Regression and Logistic Regression. In ML, classification is used to predict the outcome of a given sample when the output variable is in the form of categories. Example of classification algorithms is Naïve Bayes Classifier, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN).

On the other hand, unsupervised learning is an algorithm used for collecting population. This algorithm can describe hidden structures by exploring the unlabeled data. Example of such algorithms is K-Means, Mean Shift and K models. Meanwhile, reinforcement learning is algorithms that continuously train data via trial and error method to make specific decisions. This learning method applies to some cases with trial and error search and delayed reward [13]. In order to decide on the best decision, this method will try to apprehend the best possible knowledge by analysing sample data that had been trained before. Example of reinforcement learning algorithms includes Markov Decision Process and Q Learning.

### *C. Personality Prediction System from Facebook user*

For many years, Facebook has been using Personality Prediction Systems that can predict a user's personality automatically from their Facebook functions [6]. Facebook uses the Big Five Personality model that accurately predicts a user's personality based on someone's personality traits. Several traits can be discovered using this model such as extraversion, conscientiousness, neuroticism, agreeableness and openness. In this study, the researchers used two collections of datasets to predict the users' personality. The first dataset is samples data from the myPersonality project, and the second dataset is data that was generated manually. In the pre-processing stage, the texts written in the English language are corrected before it goes through to the next stage. Pre-processing steps consist of removing URLs, symbols, names, spaces, lowering case, stemming, and removing stop words. For data in Bahasa Melayu language, slang words or non-standard words are manually replaced in a different pre-processing stage before we translated the texts to English.

For the classification process in this study, various series of tests were conducted using deep learning and traditional machine learning algorithms for predicting the personality type of candidates for a particular job position to achieve optimum accuracy. Traditional machine learning algorithms used include Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA). Meanwhile, deep learning implementations used four architectures, namely Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and 1-Dimensional Convolutional Neural Network (CNN 1D). Results of experiments on traditional machine learning algorithms proved that in myPersonality dataset, the LDA algorithm has the most significant degree of average accuracy. Other than that, the SVM algorithm has the highest average accuracy in a manually gathered dataset (although the difference with other algorithms is not significant).

Meanwhile, the results of experiments on deep learning algorithms proved that MLP architecture has the highest average accuracy in myPersonality dataset and LSTM+CNN 1D architectures have the highest accuracy in a manual gathered dataset. In conclusion, we can improve the accuracy of datasets by using deep learning algorithms, even for traits with relatively low accuracy. This happens because, in this study, only a small number of dataset is used.

### *D. Personality Traits from Tweets using Word Embedding and Supervised Learning*

In this study, we used Twitter as a source to derive personality traits. This social media platform is a rich source of textual data, and users' behaviour, a platform people use to reflect many aspects of life, including personality. People widely share their feelings, moods, and opinions that provide a rich and informative collection of personal data that could be used for a variety of purposes [10]. Other than that, there is a recent work that constructed a questionnaire which is called Big Five Inventory (BFI) personality test for the personality traits. It consists of 44 short phrases with a five-level Likert scale, and can accurately measure the five personality traits plus six underlying facets for each trait. Then, 26 panellists were asked to share their Twitter handles and to answer the questionnaire. The pre-processing stage includes URL removal, mention removal and hashtag removal that consists of textual features created by users. Aside from that, we also removed retweets without additional content. Then, they separately fed each tweet vector to the trained model to obtain a prediction, and average all the values to compute the final personality trait score.

To derive the best performing predictive model, researchers explored different ML algorithms and performances. The ML algorithms are evaluated based on the training set through minimising the mean squared error as their loss function. They also compared the learning model

(SVM) with two baseline algorithms (Linear Regression and LASSO). These baseline algorithms are used in state-of-the-art approaches for personality prediction. The result showed that SVM classifier was able to predict the personality of Twitter users with a certain degree of accuracy, and achieve lower mean squared error. Linear Regression and LASSO models that were trained with lack of discriminative power and tend to predict personality values that are close to the average score in the myPersonality Gold Standard data.

### *E. Machine Learning Approach to Personality Type Prediction based on MBTI*

Myers–Briggs Type Indicator® (MBTI) combines 16 different personality types in four dimensions. These basic dimensions describe the preferences of an individual. The four dimensions which are also known as basic meta-programmers are Extroversion–Introversion (E–I), Sensation–Intuition (S–N), Thinking–Feeling (T–F), and Judgment–Perception (J–P). There are two types of personality for each dimension. This study predicted the personality type of a person based on the MBTI [11]. In the pre-processing stage, they collected data from an Internet forum and removed the MBTI types by using NLTK. Then, we transformed bent forms of words into their root words; a process is known as text lemmatised. Then, we categorised 16 classes of personality types into four binary classes (dimensions). Each of these binary classes represents an aspect of personality according to the MBTI personality model.

After the pre-processing stage, we created the Gradient Boosting Model. In this stage, we split the data into training and testing datasets after the MBTI type indicators were trained individually. We used training data to fit the model while testing data for predictions. Then, they used another existing method which is a recurrent neural network to determine the accuracy of the prediction. Based on the comparison, XGBoost that is based on Gradient Boosting classifier showed better accuracy than the recurrent neural network.

Before we build the new approach, we need to consider the existing systems that have been implemented to ensure that our new approach is better and constructed correctly. We made the comparison based on the personality model and method implemented in the existing systems. Table I shows a comparison of existing approaches.

TABLE I. PREVIOUS STUDIES ON PERSONALITY PREDICTION USING MACHINE LEARNING

| Studies | Personality Model | Method |
|---|---|---|
| Tandera et al., 2017 [6] | Big Five Personality Model | Traditional machine learning, Deep learning |
| Carducci et al., 2018 [10] | Big Five Personality Model | SVM Classifier, Linear Regression, LASSO |
| Amirhosseini and Kazemian, 2020 [11] | Myers–Briggs Type Indicator® (MBTI) | NLTK, XGBoost |

Based on the comparison of existing systems, we can improve the new approach to achieve more accurate data and better personality results. Also, increasing size of the dataset could potentially give a more precise prediction. In this research, we used tweets from Twitter social media extracted from Kaggle repository as our dataset.

### *F. Random Forest Classifier*

Random Forest Classifier also was known as ensemble algorithm is a supervised learning algorithm [18] that combines the same or different kind of more than one algorithm for classifying object [17]. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. A random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of observations, splitting nodes in each tree considering a limited number of features and merging them to get a more accurate and stable prediction [19]. The average prediction of each tree is the final predictions of the random forest [20]. Advantage and disadvantage of Random Forest Classifier include:

Advantages

- It is a robust method that consists of many decision trees, making it highly accurate.
- There is no overfitting problem. It will cancel out biases by taking the average of all predictions.
- The algorithm can be used in both classification and regression problems.
- This algorithm handles missing value in two ways: using median value and computing proximity-weighted average of missing value.

Disadvantages

- This algorithm requires multiple decision trees resulting to slow prediction generation process. This is because the same given input needs to be predicted and voted for all the trees in the forest, making the process time-consuming.
- The prediction model is challenging to interpret because it is hard to make a decision based on the path in the tree compared to the whole decision tree.

## III. MATERIALS AND METHODS

This section discusses the method used for developing Intelligent Personality Prediction using Machine Learning. This methodology is used to help in structuring the model development process.

### *A. Model Development*

*1) Dataset:* We collected the data from Kaggle repository (https://www.kaggle.com/datasnaek/mbti-type). In this study, the dataset contains over 8675 rows of data with two columns, as shown in Fig. 3. In each row, the data held a person's:

- Type: The person's four letters MBTI code/type

- Posts: Each of the last 50 things the person posted on Twitter. Each entry is separated by "|||" (3 pipe character).)).

We collected the dataset in 2017 from users of an online forum,personalitycafe.com (https://www.personalitycafe.com /forums/myers-briggs-forum.49/). We conducted the data collection in two phases. In the first phase, the users answered a set of questionnaire that sorts them based on their MBTI type. In the second phase, users were allowed to chat publicly with other users in the forum. The chatting sessions allowed more personality type data to be generated based on MBTI type.

*2) Exploratory data analysis:* We conducted exploratory data analysis was to get visual representation for further investigation through a violin plot printing. The number of words per comment was examined to obtain the intuitive idea of sentence structure for each personality, as shown in Fig. 4.

After that, seven additional features were created since there are currently only two features in the dataset, namely Type and Posts. The additional features are as below:

- words per comment,
- ellipsis per comment,
- links per comment,
- music per comment,
- question marks per comment,
- images per comment,
- exclamation marks per comment.

For every feature, the average number of words, punctuation, etc., are calculated. After we added these features, we analysed the Pearson correlation between words per comment and ellipses per comment for overall set of data to see how the raw data looks like and to see how the features distinguish between the four MBTI types as shown in Fig. 5. In this step, we used 'Seaborn' which is a Python data visualisation library and 'Matplotlib' which is a Python 2D plotting library for data visualisation and correlation of the MBTI personality types.

From Fig. 5, we can see that there is a high correlation between words per comment and ellipses per comment. 69% of the words are correlated with an ellipsis. To observe which personality type has the highest correlation, we charted joint plot and pair plot on the correlation variables for the different types of personality in comparison to the words per comments and ellipses per comment as shown in Table II. Fig. 6, 7, 8, and 9 shows the relationship between ellipsis per comment and words per comment for ISTP, ISTJ, ISFP, and ISFJ personality type. Meanwhile, Fig. 10, 11, 12 and 13 represents the relationship between ellipsis per comment and words per comment for INTP, INTJ, INFP, and INFJ personality type. And lastly, Fig. 14, 15, 16, and 17 shows the relationship between ellipsis per comment and words per comment for ENTJ, ENTP, ENFP, and ENFJ personality type.

```
   type                                               posts
0  INFJ  'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1  ENTP  'I'm finding the lack of me in these posts ver...
2  INTP  'Good one  _____   https://www.youtube.com/wat...
3  INTJ  'Dear INTP,   I enjoyed our conversation the o...
4  ENTJ  'You're fired.|||That's another silly misconce...
5  INTJ  '18/37 @.@|||Science  is not perfect. No scien...
6  INFJ  'No, I can't draw on my own nails (haha). Thos...
7  INTJ  'I tend to build up a collection of things on ...
8  INFJ  I'm not sure, that's a good question. The dist...
9  INTP  'https://www.youtube.com/watch?v=w8-egj0y8Qs||...
****************************************
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8675 entries, 0 to 8674
Data columns (total 2 columns):
type     8675 non-null object
posts    8675 non-null object
dtypes: object(2)
memory usage: 135.7+ KB
None
```

Fig. 3. MBTI Personality Type Dataset.

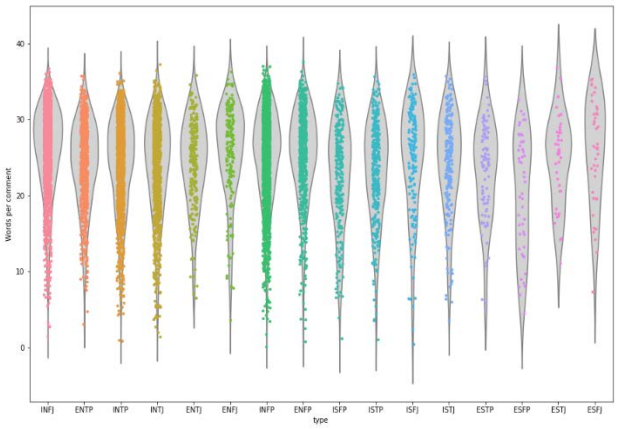


Fig. 4. Words Per Comment for each Personality Type.

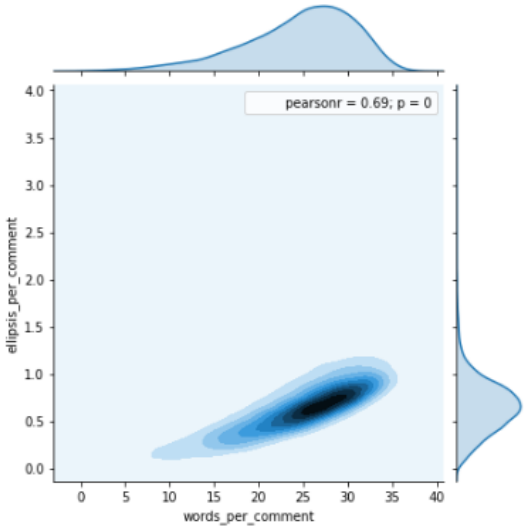


Fig. 5. Pearson Correlation.

Each of the personality comes with the results of Pearson correlation (pearsonr = 0.73). For Fig. 8 and 12, 72% of the words are correlated with an ellipsis. Other than that, for Fig. 10, 13, and 16, 64% of the words are correlated with an ellipsis. Meanwhile, for Fig. 11 and 14, 74% of the words are correlated with an ellipsis.

TABLE II. PEARSON CORRELATION FOR WORDS PER COMMENT VS ELLIPSES PER COMMENT FOR EACH MBTI PERSONALITY TYPE

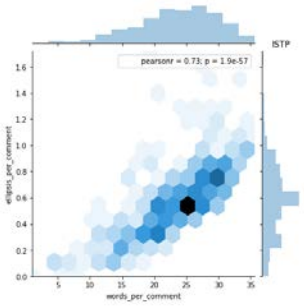


Fig. 6. ISTP Personality Type.

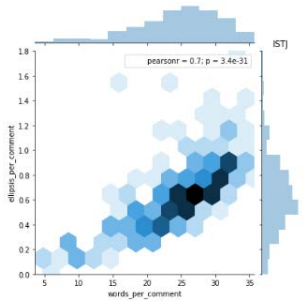


Fig. 7. ISTJ Personality Type.

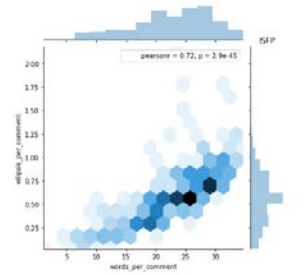


Fig. 8. ISFP Personality Type.

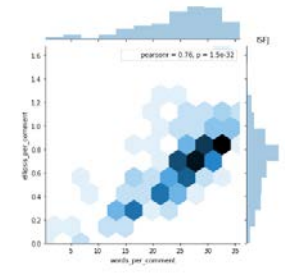


Fig. 9. ISFJ Personality Type.

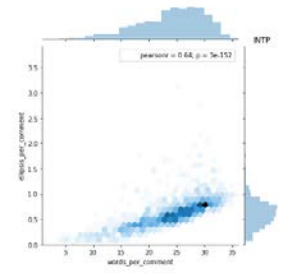


Fig. 10. INTP Personality Type.

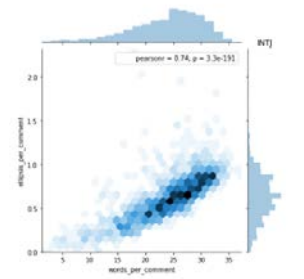


Fig. 11. INTJ Personality Type.

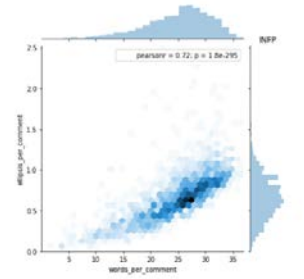


Fig. 12. INFP Personality Type.

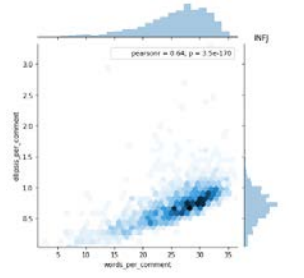


Fig. 13. INFJ Personality Type.

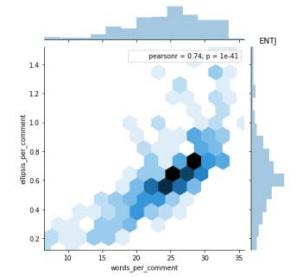


Fig. 14. ENTJ Personality Type.

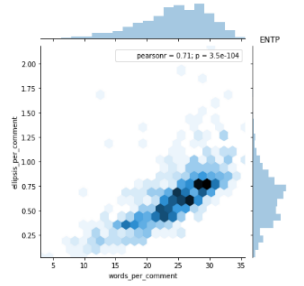


Fig. 15. ENTP Personality Type.

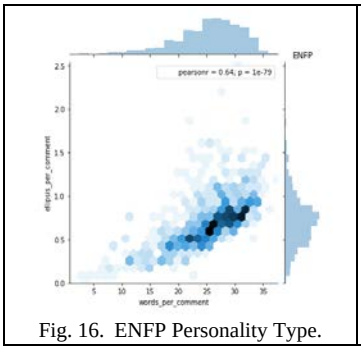


Fig. 16. ENFP Personality Type.

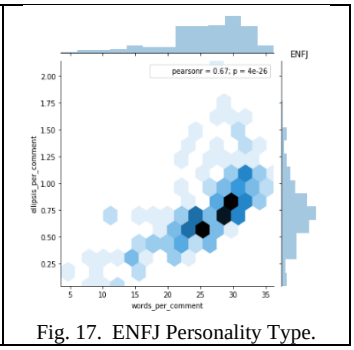


Fig. 17. ENFJ Personality Type.

From Fig. 18, the top three highest correlation values for the ellipses per comment and words per comment are:

- INFJ – The advocate - Introversion Intuition Feeling Judging
- INTP - The Thinker - Introversion Intuition Thinking Perceiving
- ENFP - The Inspirer - Extroverted Intuition Feeling Perceiving

From this exploratory data phase, each MBTI type has a different correlation between ellipses per comment and words per comments. The correlation determines how closely each feature is affected by another feature. INFJ, INTP and ENFP recorded the highest correlation, which is an excellent sign to train the data and build machine learning models.

*3) Data pre-processing:* To get further insight on the dataset, we created four new columns that divided the respondents based on the four dimensions of MBTI namely Extroversion–Introversion (E–I), Sensation–Intuition (S–N), Thinking–Feeling (T–F), and Judgment–Perception (J–P). The process is to improve the accuracy of the results.

Furthermore, we also used word2vec technique in this pre-processing step. Word2vec is an algorithm to construct vector representations of words, also known as word embedding. In this paper, we converted textual data into numeric signals. For example:

- I = 0, E = 1
- N = 0, S = 1
- T = 0, F = 1
- J = 0, P = 1

*4) Dataset splitting:* To test the model's accuracy, we split the dataset into two parts which were training dataset and testing dataset. We used 90% of data for training, and 10% for testing and keeping random state five using sci-kit learn's internal module train_test_split (). The testing dataset is a set of unseen data that was used only to access the performance of a fully specified classifier.

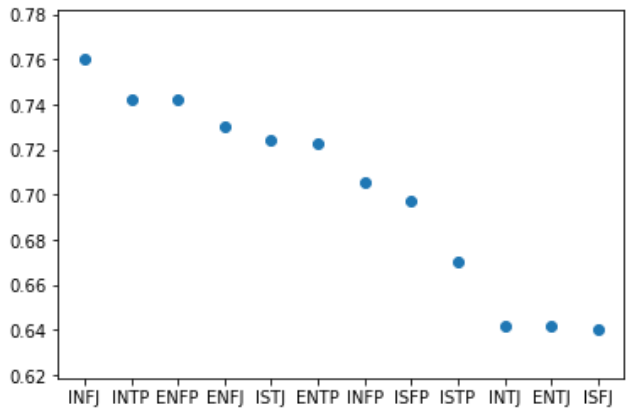


Fig. 18. Pair Plot of Pearson Correlation For words Per Comment vs Ellipses Per Comment.

*5) Model building:* To build the Random Forest, Logistic Regression, KNN Neighbor and Support Vector Machine (SVM) models, we used Numpy and sklearn. We use train_test_split function from sklearn library to split the data into training and testing datasets while the MBTI type indicators were trained individually. In total, we used 90% of the data for the training set (data fitting), and 10% for testing (making a prediction). We first remove all columns irrelevant to our features. From there, we can see that the Random Forest algorithm was able to classify all respondents (100%) to the right types. Then, we do a deeper dive into our model to get a better perspective on our prediction by performing four machine learning algorithms with Extroversion–Introversion (E–I) column, Sensation–Intuition (S–N) column, Thinking–Feeling (T–F) column, and Judgment–Perception (J–P) column.

*6) Comparing the accuracy of machine learning models:* In this step, the accuracy of the Random Forest and three other models namely Linear Regression, KNN neighbor and Support Vector Machine (SVM) were evaluated using the testing dataset.

*7) Evaluating results:* Evaluation of the results helps in finding the best model that represents the data. The result of this evaluation is presented in Section 4.

## IV. Results and Discussion

This section discusses the results of the experiment conducted. We did several experiments to obtain the most significant model for predicting MBTI personality types. Firstly, we determine the arrangement of the MBTI personality types by calculating words per comment in the dataset. We added several features in this experiment since the original dataset only has two features. We analysed these features by calculating the average of each feature, namely average words per comment and average ellipses per comment. After that, we measured the Pearson correlation coefficient to know the strength between variables and relationships. Since there is a large correlation (69%) between words per comment and ellipses per comment, we chose this variable to train the machine learning model.

From Pearson correlation conducted, it is evident that INFJ, INTP, and ENFP personality types have the highest correlation between words per comment and ellipses per comment. Next, data pre-processing using word2vec technique was done to make the dataset more organised and easy to understand. Lastly, we use train_test_split function from sklearn library to split the data into training and testing datasets while the MBTI type indicators were trained individually. We used training data to fit the model and testing data for prediction. The last step is we develop four machine learning models, and we obtained the accuracy of each machine learning model for every MBTI personality type.

Table III shows that the Random Forest model has better accuracy (100%) in all four dimensions of MBTI personality types compared to other machine learning models. Accuracy of the Random Forest model is considerably higher than the Support Vector Machine (SVM) model for Intuition (I)–Sensation (S) and Introversion (I)–Extroversion (E) categories, while for Sensation–Intuition (S–N) category, the accuracy is a little bit better. Accuracy of SVM for Judgment (J)–Perception (P) is considerably worse than the Random Forest model. Thus, the overall performance of the Random Forest model is better than the three other machine learning models for this dataset.

TABLE III. Results

| Model | E vs I | S vs N | F vs T | J vs P | Overall |
|---|---|---|---|---|---|
| Random Forest | 100% | 100% | 100% | 100% | 100% |
| Logistic Regression | 77.11% | 86.03% | 63.35% | 60.37% | 23.35% |
| KNN neighbor | 83.66% | 88.11% | 77.64% | 77.74% | 40.62% |
| Support Vector Machine (SVM) | 77.16% | 86.03% | 56.54% | 47.16% | 16.94% |

## V. Conclusions and Future Works

In conclusion, this research could predict personality by using social media data, and the best model of a machine learning algorithm, which are the Random Forest machine learning algorithm. With that, this will significantly benefit companies because they can analyse their candidates' social media accounts before they choose the right employees.

### A. Limitations

This research only studied people with particular social media, namely Twitter. The are other social media platforms that could give significant data, thus improve the prediction model. In addition, this research only focused on the prediction of strengths and weaknesses in terms of personality. It is essential to consider a technical position on the team when we are creating teams to fight crime, develop unique software, or play sports. Aside from this, we also need to explore people's soft skills. It is also essential to consider other factors, namely, mindset and personality. In short, this is just the first step in creating a personality type model based on MBTI personality assessment from social media comment data.

This research trains the model based on a large number of tweets, and it is not easy to collect such a massive database for this process. By improvising this research, future research can use a small number of tweets for both training and testing to examine the performance of the method. Finally, this research only used English data. To improve this, we recommend future research to study on multiple social media platforms or different cultures. They can use various data sources to get more insights and exciting finding, using machine learning approaches.

### *B. Future works*

In the future, we plan to collect and build more datasets to get a more accurate result. We also plan to use XGBoost algorithm and deep learning algorithm, their architectures, and other processes to improve this prediction system. XGBoost algorithm which optimised distributed gradient boosting machines is scalable and well-known for its excellent performance in terms of computational speed. This algorithm can push the limits of computing power for boosted trees algorithms. Other than that, deep learning is also a suitable candidate to address this challenge as it can generate new features from a limited series of features located in training datasets. Due to this, the method requires less time to analyse big data.

## ACKNOWLEDGEMENT

This work was funded by Universiti Malaysia Pahang [RDU190396] and Universiti Malaysia Kelantan via UMK Fund [R/FUND/A0100/01850A/001/2020/00816].

Conflicts of Interest: The authors declare no conflicts of interest.

## REFERENCES

[1] X. Teng and Y. Gong, “Research on Application of Machine Learning in Data Mining,” IOP Conf. Ser. Mater. Sci. Eng., vol. 392, no. 6, 2018.

[2] F. Ahmed, P. Campbell, A. Jaffar, S. Alkobaisi, and J. Campbell, “Learning & Personality Types: A Case Study of a Software Design Course,” J. Inf. Technol. Educ. Innov. Pract., vol. 9, no. January, pp. 237–252, 2010.

[3] B. de Raad and B. Mlačić, “Big Five Factor Model, Theory and Structure,” Int. Encycl. Soc. Behav. Sci. Second Ed., no. December, pp. 559–566, 2015.

[4] T. L. C. Yoong, N. R. Ngatirin, and Z. Zainol, “Personality prediction based on social media using decision tree algorithm,” Pertanika J. Sci. Technol., vol. 25, no. S4, pp. 237–248, 2017.

[5] N. R. Ngatirin, Z. Zainol, and T. L. C. Yoong, “A comparative study of different classifiers for automatic personality prediction,” Proc. - 6th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2016, pp. 435–440, 2017.

[6] T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, “Personality Prediction System from Facebook Users,” Procedia Comput. Sci., vol. 116, pp. 604–611, 2017.

[7] W. Rc, Y. Munas, K. Cs, F. Ta, and Vithana N, “Personality Based E-Recruitment System,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, 2017.

[8] S. Sharma, “Predicting Employability from User Personality using Ensemble Modelling,” no. July, p. 37, 2015.

[9] B. Y. Pratama and R. Sarno, “Personality classification based on Twitter text using Naive Bayes, KNN and SVM,” Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015, no. November, pp. 170–174, 2016.

[10] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, “TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning,” Inf., vol. 9, no. 5, pp. 1–20, 2018.

[11] M. H. Amirhosseini and H. Kazemian, “Machine Learning Approach to Personality Type Prediction Based on the Myers – Briggs Type Indicator ®,” 2020.

[12] N. Dhanda, S. S. Datta, and M. Dhanda, “Machine Learning Algorithms,” no. June 2016, pp. 210–233, 2019.

[13] G. M., T. A., S. C., and R. P., “Soybean Under Water Deficit: Physiological and Yield Responses,” A Compr. Surv. Int. Soybean Res. - Genet. Physiol. Agron. Nitrogen Relationships, 2013.

[14] T. Varvel, S. G. Adams, and S. J. Pridie, “A study of the effect of the myers-briggs type indicator on team effectiveness,” ASEE Annu. Conf. Proc., pp. 9525–9533, 2003.

[15] Z. Poursafar, N. Rama Devi, and L. L. R. Rodrigues, “Evaluation of Myers-Briggs Personality Traits in Offices and Its Effects on Productivity of Employees: an Empirical Study,” Res. Artic. Int J Cur Res Rev, vol. 7, no. 21, pp. 53–58, 2015.

[16] S. D. Mallari, “Myers-Briggs Type Indicator (MBTI) Personality Profiling and General Weighted Average (GWA) of Nursing Students.,” Online Submiss., no. October, pp. 1–11, 2017.

[17] “Chapter 5: Random Forest Classifier - Machine Learning 101 - Medium.” [Online]. Available: https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1. [Accessed: 02-May-2020].

[18] “Random Forests Classifiers in Python - DataCamp.” [Online]. Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python#features. [Accessed: 02-May-2020].

[19] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[20] “An Implementation and Explanation of the Random Forest in Python.” [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76. [Accessed: 02-May-2020].