Karbala International Journal of Modern Science

Volume 8 | Issue 4 Article 5

Predicting Users' Personality on Social Media: A Comparative Study of Different Machine Learning Techniques

Ali Saadi Al-Fallooji
Department of Software/ College of Information Technology/ University of Babylon, alisadee.sw.msc@student.uobabylon.edu.iq

Ahmed Al-Azawei
Department of Software/ College of Information Technology/ University of Babylon

Recommended Citation
Al-Fallooji, Ali Saadi and Al-Azawei, Ahmed (2022) "Predicting Users' Personality on Social Media: A Comparative Study of Different Machine Learning Techniques," *Karbala International Journal of Modern Science*: Vol. 8 : Iss. 4 , Article 5.
Available at: https://doi.org/10.33640/2405-609X.3262

Predicting Users' Personality on Social Media: A Comparative Study of Different Machine Learning Techniques

Abstract

The use of social media sites (SMSs) becomes ubiquitous worldwide as the number of users is noticeably increasing. This has led to exploiting such sites by market, business, and educational companies to deliver content that meets users' personal needs. However, this requires identifying users' personalities to respond to their individual preferences. This research aims at (1) analyzing users' posts on SMSs to predict their personality based on the Meyers-Briggs Type Indicator (MBTI) model, (2) comparing the performance accuracy of different preprocessing and data mining techniques, and (3) improving the prediction accuracy of users' personality types. The used dataset includes 8668 records in which each raw contains fifty posts. Three data mining techniques are applied namely, support vector machine (SVM), logistic regression (LR), and lightGBM. The findings suggest that lightGBM with the application of stemming, lemmatization, and grid search optimization as well as removing stop-words outperformed other techniques. The prediction accuracies for the four personality dimensions namely, Introversion-Extroversion (I-E), Intuition-Sensing (N-S), Feeling-Thinking (F-T), and Judging-Perceiving (J-P) are 100.0%. The research outcomes are promising as the four dimensions of MBTI have been identified effectively. Such outcomes are also compared with earlier research on personality prediction. This study can help SMSs providers, businesses, and educational institutions adapt their online sites based on users' posts, tweets, and comments that can be used to predict their personality behavior.

Keywords

Social media sites (SMSs), Users' personality, Myers-Briggs Type Indicator (MBTI), machine learning

RESEARCH PAPER

Predicting Users' Personality on Social Media: A Comparative Study of Different Machine Learning Techniques

Ali S. Al-Fallooji*, Ahmed Al-Azawei

Department of Software, College of Information Technology, University of Babylon, Iraq

Abstract

The use of social media sites (SMSs) becomes ubiquitous worldwide as the number of users is noticeably increasing. This has led to exploiting such sites by market, business, and educational companies to deliver content that meets users' personal needs. However, this requires identifying users' personalities to respond to their individual preferences. This research aims at (1) analyzing users' posts on SMSs to predict their personality based on the Meyers-Briggs Type Indicator (MBTI) model, (2) comparing the performance accuracy of different preprocessing and data mining techniques, and (3) improving the prediction accuracy of users' personality types. The used dataset includes 8668 records in which each raw contains fifty posts. Three data mining techniques are applied namely, support vector machine (SVM), logistic regression (LR), and lightGBM. The findings suggest that lightGBM with the application of stemming, lemmatization, and grid search optimization as well as removing stop-words outperformed other techniques. The prediction accuracies for the four personality dimensions namely, Introversion-Extroversion (I-E), Intuition-Sensing (N−S), Feeling-Thinking (F-T), and Judging-Perceiving (J-P) are 100.0%. The research outcomes are promising as the four dimensions of MBTI have been identified effectively. Such outcomes are also compared with earlier research on personality prediction. This study can help SMSs providers, businesses, and educational institutions adapt their online sites based on users' posts, tweets, and comments that can be used to predict their personality behavior.

Keywords: Social media sites (SMSs), Users' personality, Myers-Briggs type indicator (MBTI), Machine learning

1. Introduction

Social media sites (SMSs) are used by people to build friendships with others who have similar interests in personal or vocational activities. SMSs have become a part of people's life because they use such sites to check, share, or like their friends' posts and perspectives [1]. In the last few years, there has been a huge surge in the amount of information that people have, especially in the form of textual data. People can send text messages on a lot of different websites such as Twitter, Facebook, Instagram, YouTube, and TikTok. Each day, the average time spent by a person on SMSs is between two and three hours. This comes with the increase in the number of SMSs' users which is two billion for YouTube, two billion for WhatsApp, 1.3 billion for Facebook Messenger, 1.3 billion for Twitter, and 1.2 billion for WeChat [2]. People utilize social media to express themselves on themes such as life and family, psychology, finance, society and environment, and politics [3]. A prior study found a high association between user personality behavior and his/her behavior on social media [3]. Some of the applications that can benefit from personal information are recruitment systems, personal consulting systems, and online marketing [4]. Thus, personalizing SMSs should highly rely on users' personality behavior.

A personality is a group of things that make a person unique from other people such as his/her characteristics, thoughts, feelings, and behaviors [5].

Received 2 April 2022; revised 5 August 2022; accepted 8 August 2022.
Available online 10 November 2022

* Corresponding author at:
E-mail address: alisadee.sw.msc@student.uobabylon.edu.iq (A.S. Al-Fallooji).

https://doi.org/10.33640/2405-609X.3262

It is a term used in psychology that talks about how personality and psychological disorders affect job performance and job satisfaction [6]. Behavior modification and modulation are the main areas that provide people with a reason to interact with each other and have balanced relationships. However, it can be hard to figure out which psychological types a person belongs to because they vary. Many different models of personality have been proposed in the literature of psychology such as the Myers-Briggs Type Indicator (MBTI) [7], Dominance influence steadiness conscientiousness (DISC) [8], Strength Finder [9], and Big Five Personality Traits (BFT) [10]. In this study, the MBTI model is used because it is the most commonly used theory in adaptive technology [10,11].

MBTI helps people understand how they work and learn. Understanding peoples' personality represents a successful way to build relationships, be more positive, and do well [12]. MBTI includes four dimensions which are: Introversion (I) vs. Extroversion (E), Intuition (N) vs. Sensing (S), Feeling (F) vs. Thinking (T), and Judging (J) vs. Perceiving (P). The four pairs combine dimensions into 16 different types of personality, and they can be coded as a set of sixteen different types of personality. These are ESTJ, ISTJ, ESTP, ISTP, INFP, ISFP, INTJ, ISFJ, INTP, INFJ, ENFP, ENTJ, ESFJ, ENFJ, ENTP, and ESFP. The model has been widely used in business for several different reasons such as employee profiling and/or promotion [13].

In personalizing information systems research, a lot of different ways can be used to figure out how people think. Such methods can be categorized into two groups namely, explicit and implicit methods. The formal is to collect direct feedback from users such as filling out a questionnaire. This method, on the other hand, has a lot of issues. Respondents may not be 100% honest with their answers [14]. Furthermore, there is a chance that some questions are not answered. Users may also understand or interpret questions wrongly. Finally, surveys may cause people fatigue if they are too long [15]. To avoid such drawbacks, a new direction called the implicit approach was proposed in the literature to identify users' personalities based on analyzing their own comments, posts, and tweets. In particular, previous research showed low to moderate accuracy in predicting users' personalities, particularly for Judging-Perceiving (J/P) and Sensing-Intuition (S/N) dichotomies of MBTI [16−18]. Thus, the prediction of MBTI still needs further research.

Accordingly, it is important to build an effective model that can predict users' personalities based on their texts or posts on SMSs. The progress made in the domain of natural language processing (NLP) can be exploited. NLP has made it possible for computers to interpret words or sentences written in human language [19,20]. Part-of-speech (noun, verb, and adjective) and grammar structures are used in NLP [21]. In some ways of implicit analysis, users' profile pictures and other pictures that they share can also be used to figure out what kind of person they are and make information systems more personalized [22]. This can also be achieved based on a user's personal information in his/her SMSs account [23]. This research, therefore, aims to analyze users' comments on SMSs to predict users' personalities. This can help personalize such platforms according to users' needs and preferences. In order to achieve this aim, a classification model is built to predict personality behavior based on MBTI and the following objectives are covered:

1. Analyzing users' posts on SMSs.
2. Building a classification model to predict personality behavior.
3. Comparing the prediction ability of different machine learning methods.
4. Enhancing the prediction accuracy of the implemented classifiers.

In this research, the overall outcomes outperform all earlier literature to be the key contribution of this work. Other research contributions are: (1) extending earlier literature on personalizing information systems implicitly, (2) enhancing the accuracy of the classification model by following a rigorous preprocessing procedure and by applying the grid search optimizer, (3) comparing the execution time of different classifiers in which this was neglected in previous research, and finally (4) comparing the research findings of different machine learning techniques.

The rest of the paper is organized as follows. The relevant literature is reviewed in Section 2. Section 3 presents the theoretical foundation and the key concepts of the proposed model. Section 4 shows the research findings. Section 5 discusses the research outcomes alongside the previous research. Section 6 shows the theoretical implications. The practical implications are explained in Section 7. Section 8 concludes the key outcomes of this research, its limitations, and the possible future research directions.

2. Literature review

Previous studies attempt to predict users' personalities following several different approaches and techniques. After a comprehensive review that has been conducted in this research, it was found that three key directions that have been adopted in related literature. The most dominant direction is based on using machine learning techniques which represents the main focus of this present research. Other studies implemented deep learning techniques to identify users' individual personalities on SMSs. Another piece of literature considers multiple accounts of a user on different social media platforms to predict his/her personality.

2.1. Machine learning

Previous literature exploits the advantages of NLP to predict users' personalities on SMSs. In [17], the MBTI model was used to analyze the performance of different classifiers in which users' personalities were predicted based on their online text. Two supervised algorithms were applied which are Bidirectional Encoder Representations from Transformers (BERT) and Naive Bayes (NB). The best accuracy achieved was 61% based on the NB algorithm. In another study [24], MBTI dimensions were predicted on social media. The employed methods were support vector machine (SVM), NB, random forest (RF), and logistic regression (LR). The best accuracy was 65.4%. Another research study suggested a new MBTI dataset based on the Reddit social media network for personality prediction [18]. The XGBoost technique was employed in which the highest accuracy obtained was 76.1%. In [16], the prediction of the Judging-Perceiving (J/P) dichotomy was low. This was because it is difficult to anticipate the J/P dichotomy as it involves looking at people's actions and behaviors. The J/P binary is not related to the number of posts or comments. There was also difficulty in predicting Sensing-Intuition (S/N) dichotomy because the Intuitive type dominated in the dataset and this, in turn, led to an unbalanced dataset [17].

This paper, therefore, aimed to solve some of the above-mentioned problems in previous literature since the achieved accuracy was ranging between 60% and 85%. Table 1 summarizes earlier research. It encompasses the dataset used in previous literature to predict users' personality on SMSs, the key methods used in the data preprocessing, the classification techniques, and the highest accuracy achieved for each dimension.

2.2. Deep neural networks

In terms of using more advanced prediction techniques, deep learning has been widely applied to enhance the performance accuracy of predicting users' personalities. Earlier literature that

Table 1. A summary of the related work.

Reference	Dataset	Method	Feature Extraction	
[25]	Kaggle MBTI	Naïve Bayes & SVM	TF-IDF + LIWC	
	Lemmatization	**Stemming**	**Stop-words**	
	✓	✓	X	
	Accuracy			
	IE	**FT**	**SN**	**JP**
	77.0	86.2	77.9	62.3
[26]	Twitter	Logistic Regression	LIWC	
	Lemmatization	**Stemming**	**Stop-words**	
	✓	✓	X	
	Accuracy			
	IE	**FT**	**SN**	**JP**
	84.9	88.4	87.0	78.8
[17]	Kaggle MBTI	Naïve Bayes	TF-IDF + N-gram	
	Lemmatization	**Stemming**	**Stop-words**	
	✓	X	X	
	Accuracy			
	IE	**FT**	**SN**	**JP**
	59.0	57.0	63.0	62.0
[18]	Kaggle MBTI	XGBoost	TF-IDF	
	Lemmatization	**Stemming**	**Stop-words**	
	X	X	X	
	✓	✓	✓	
	Accuracy			
	IE	**FT**	**SN**	**JP**
	78.1	86.0	71.7	65.7
	79.0	85.9	74.1	65.4

implemented deep learning achieved an overall accuracy between 59% and 88% [27,28]. On the other hand, the application of machine learning algorithms outperformed the performance accuracy of deep learning [16,18,26]. This could be attributed to two possible reasons. The first is that deep learning techniques need huge data to learn adequately [29]. The other reason is the catastrophic forgetting of the artificial neural networks when they unexpectedly forget the previously learned information while learning new information [30].

2.3. Predicting users' personalities with multiple accounts

Regarding the problem of predicting users' personalities with multiple accounts on different social networking sites, previous studies found that the prediction accuracy varies from account to account [31–33]. This is because some platforms allow users to upload photos, add friends and post comments, whereas others have restrictions on users' activities such as allowing them to post a blog with a limited number of words/characters or a small number of photos [32,33]. However, personality analysis and prediction on social media are not only limited to users' posts and/or comments [34]. Personality can also be analyzed through profile pictures, shared items, and the pages the users follow [35]. Thus, integrating users' accounts on different social media sites may provide a better understanding of their personalities and behavior.

3. Theoretical background and the proposed model

3.1. Overview of the MBTI personality model

Figure 1 represents the four dimensions of MBTI. It is a model used to determine a person's personality type in how people perceive the world, make decisions, preferences, and strengths [36]. It is the most commonly used theory for personality categorization. The model attempts to determine four categories namely:

- Introversion (I) vs. Extroversion (E).
- Intuition (N) vs. Sensing (S).
- Feeling (F) vs. Thinking (T).
- Judging (J) vs. Perceiving (P).

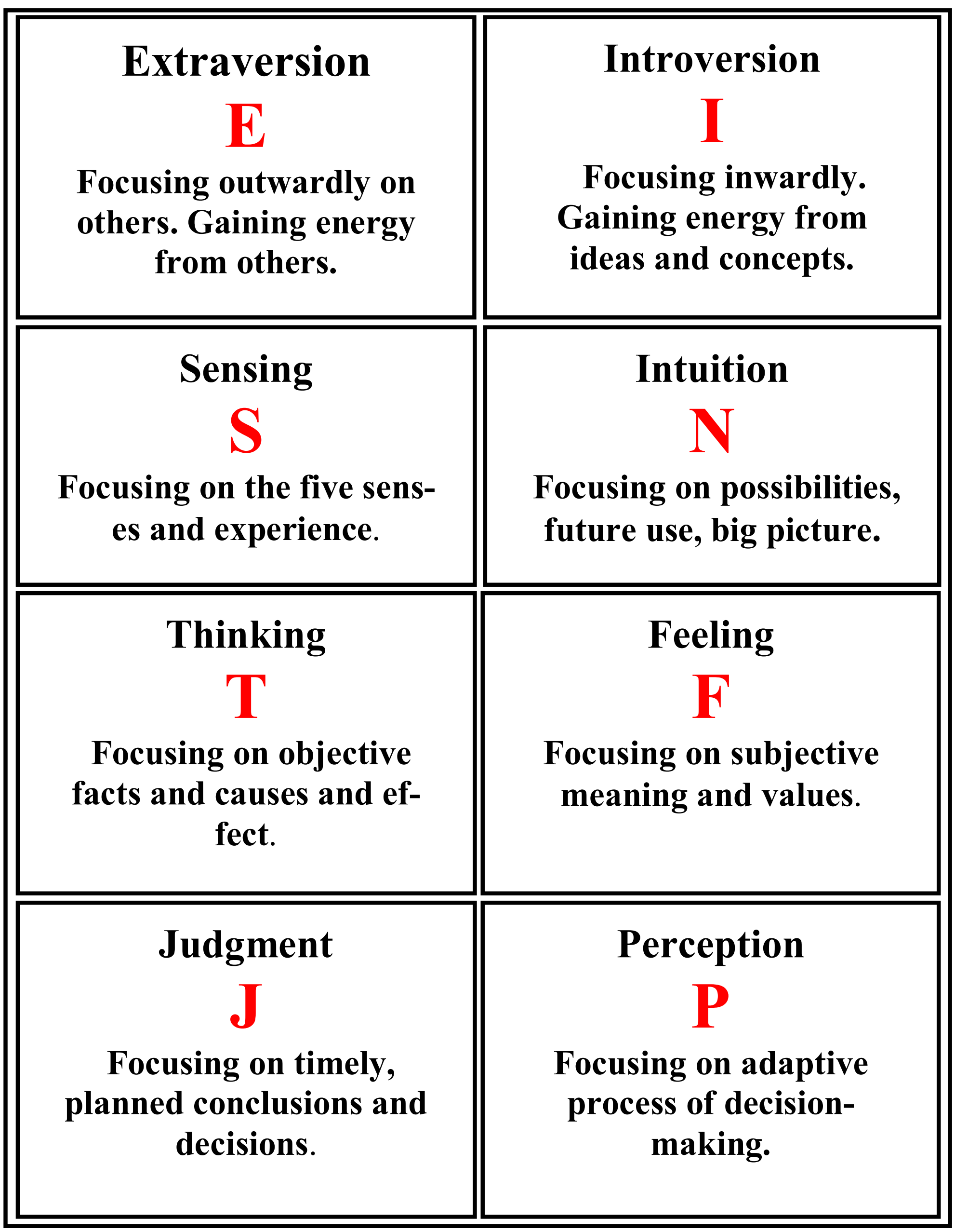

Fig. 1. The four categories of Myers–Briggs Type Indicator.

- E vs. I: Extraversion people are energized by reacting with others, taking part in activities, and are noted for responding swiftly. They are also excited about ideas, thinking, and working alone. Extraversion people usually think about what they are going to do before they do it [37].
- N vs. S: Intuition people concentrate on patterns, and future possibilities, and take pleasure in abstract thought. The main focus of Sensing people, on the other hand, is on facts and their own real-world experiences [36].
- F vs. T: Feeling people are careful about taking into account persons, feelings, and different points of view. On the other side, Thinking people are logical, highly analytical, and capable of evaluating the facts [36].
- P vs. J: Perceiving people are nimble and quick to adjust to changes in their surroundings. Judging people make goals and lists, stick to a schedule, and keep track of all needs to do [38].

To determine types of personality, each person can be classified in terms of one of the sixteen possible four-letter symbols (e.g., ESFJ, ENFP, INTP, and ISFJ). As shown in Fig. 2, each type is used to establish a unique set of behaviors for each individual. Different mindsets, orientations, and decision-making are reflected.

3.2. The proposed methodology

A dataset from the personality cafe forum is utilized to forecast individuals' personalities based on their posts and comments. The dataset is preprocessed before ap-plying feature extraction and selection algorithms. The extracted characteristics are trained using data mining until the best accuracy is achieved. The major phases of the proposed system are depicted in Fig. 3.

3.2.1. The research dataset

This study uses the personality cafe forum dataset released in 2017. This dataset can be found on Kaggle at https://www.kaggle.com/datasnaek/mbti-type. In the dataset, there are 8668 rows of data. Each row has a person's posts in which they include a mean of 1220 words [16]. The dataset had only two columns. The first is the MBTI users' type, whereas the second is what people said on the personality cafe forum. The reasons why this dataset was chosen are that (1) it is a public dataset with a large size, and (2) it is not based on microblogging. The personality cafe forum is a platform for people to converse and discuss their personalities. This research uses **Algorithm 1** to convert the class column in the original dataset into four dimensions in which each dimension contains either 0 or 1.

Algorithm 1.

Algorithm 1
Defining the algorithm - Converting the sixteen types of personality into four dimensions, each dimension is either 0 or 1.
Input - The sixteen types of personalities for the MBTI model as found in the dataset.
Output - The four dimensions of the MBTI in which each dimension is either 0 or 1.
Variables' definition -Type: It is the name of the column in the dataset that contains the sixteen types of personality. -I, N, F, and J are equal to zero, whereas E, S, T, and P are equal to one. -row: This is a parameter inside the function. - i: It is an index to reach all rows in the dataset.
A function that contains one parameter:(row) **Step 1:** Converting the first dimension (Introversion (I), Extraversion (E)) into either 0 or 1 (binary classification). Type = row['type'] if Type[i,0]='I' → 'I' = 0. else if Type[i,0]='E' → 'I' = 1. else: print ('I, E not found') **Step 2:** Converting the second dimension (Intuition (N), Sensing (S)) into either 0 or 1 (binary classification). if Type[i,1]='N' →'N' = 0. Else if Type[i,1]='S' →'N' = 1. else: print ('N, S not found') **Step 3:** Converting the third dimension (Feeling (F), Thinking (T)) into either 0 or 1 (binary classification). if Type[i,2]='F' →'F' = 0. Else if Type[i,2]='T' →'F' = 1. else: print ('T, F not found') **Step 4:** Converting the fourth dimension (Judging (J), Perceiving (P)) into either 0 or 1 (binary classification). if Type[i,3]='J' → 'J' = 0. Else if Type[i,3]='P' → 'J' = 1. else: print ('J, P not found') return ({'IE': I, 'NS': N, 'FT': F, 'JP': J})

3.2.2. Preprocessing

Different methods of text preprocessing are used in this research. Lowercasing is one of the simplest and most effective types of text preparation. It can be applied to most text mining and NLP problems and can aid in circumstances where the dataset is not particularly large [16]. Such techniques can improve the prediction accuracy of a classifier. Another preprocessing approach is stemming which refers to the process of reducing inflection in words. Stemming is a primitive heuristic procedure that chops off the ends of words in the hope of appropriately changing them into their base form [39].

Moreover, lemmatization is quite similar to stemming in its overall purpose. It eliminates inflections and maps a word to its root form. It returns the changed words to their root. Another preprocessing step is the stop-words elimination. Stop-

INTJ	INTP	ISFJ	ENTP
INFJ	ESFP	ESTJ	ENFP
ISTJ	INFP	ISFP	ESFJ
ISTP	ENFJ	ESTP	ENTJ

Fig. 2. The sixteen possible personality types of MBTI.

words refer to a collection of widely used words in a language. Examples of stop-words in English are “a", “the”, “is”, “are”, etc. The idea behind employing stop-words is to eliminate low-information terms from the text [40]. Finally, the pretreatment process includes: (1) Removing Punctuations, (2) Removing links, (3) Removing very short words, (4) Removing MBTI personality words, and (5) Removing tags from comments and posts.

Based on different preprocessing steps followed in this research, the dataset was split into eight types based on some particular preprocessing steps. For example, one type includes stop-words, whereas the other does not contain them. In another form, stemmers were used to convert all words into their stem, while the input of classifiers was without stemming in another step. Moreover, lemmatization was used to remove the additions from the words, while in another form, words were kept without lemmatization. Another form applies stop-words, stemming, and lemmatization, whereas the original dataset was also used with stop-words and without stemming and lemmatization. In this research, the preprocessing steps are structured as follows:

1. Removing tags, links, punctuations, and stop-words.
2. Splitting the words in the dataset.
3. Applying stemming and lemmatization.
4. Removing words that consist of two letters or less.
5. Removing MBTI personality words.

3.2.3. Feature extraction

This research uses feature extraction in NLTK which is the construction of vocabulary. It also calculates the weights of the features using the term frequency-inverse document frequency (TF-IDF) [40]. TF-IDF is a metric for determining how significant a term is in a text. There are three major

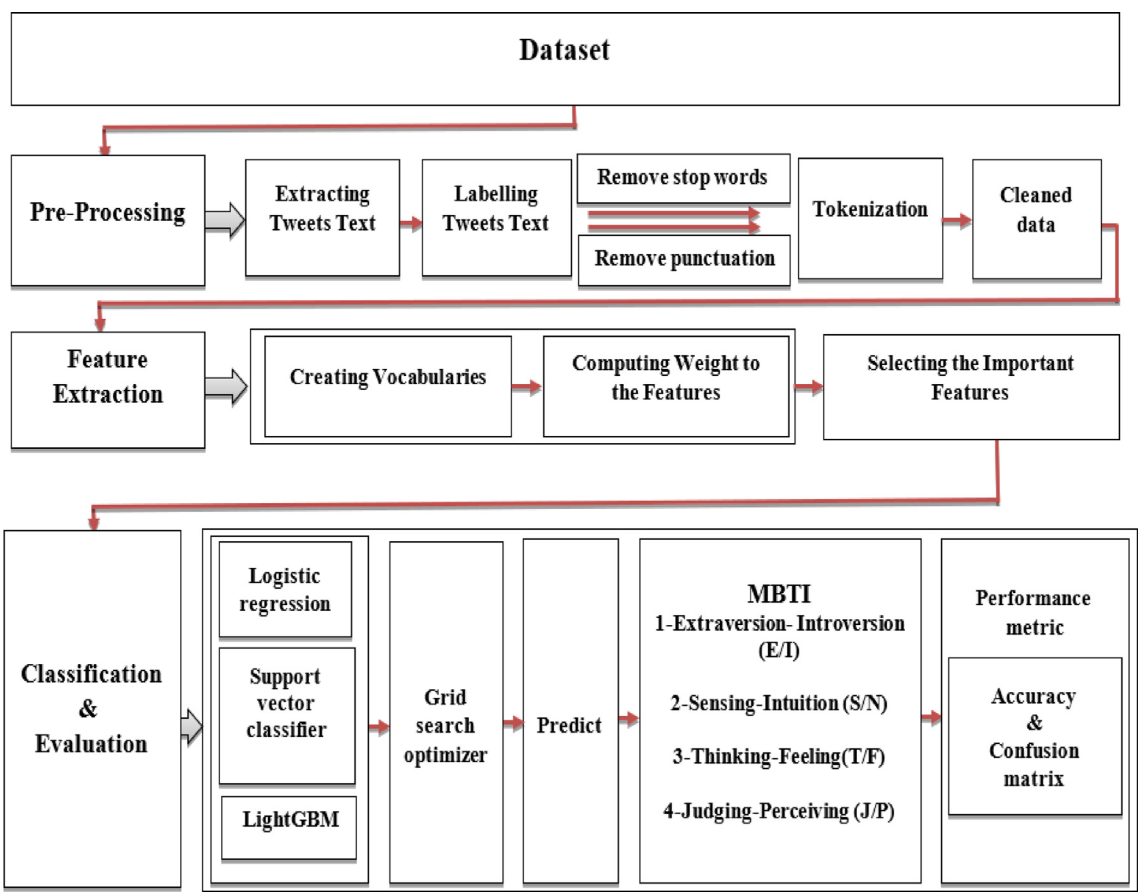

Fig. 3. The proposed methodology.

implementations for TF-IDF. These are in machine learning, information retrieval, and keyword extraction/text summarization. Machine learning Algorithms often utilize numeral data, to handle textual data or any natural language processing (NLP) errand which is a sub-field of machine learning/artificial intelligence [17]. Thus, data needs to be switched to a vector of numeral data by an operation known as vectorization. TF-IDF vectorization calculates the TF-IDF weight for every word in the dataset and then puts that information into a vector. Hence, each post or comment in the dataset has its own vector, and the vector would have a TF-IDF score for every singular word in the dataset. After obtaining these vectors, they can be used for many different aspects such as seeing if two words are similar by comparing their TF-IDF vectors. In text summarization and keyword extraction, TF-IDF is also used [40]. Because TF-IDF weights words based on how important they are, this method can be used to figure out which words are the most important and this, in turn, can help determine keywords for a dataset.

3.2.4. *Classification*

After identifying important features and filtering data, this study uses data mining to predict MBTI personality traits. Three classifiers are implemented, and their findings are compared to highlight the best accuracy that can be achieved in MBTI four dimensions prediction.

The implemented techniques were selected because of their popularity in earlier research and to compare the findings of this study with the previous literature [16,39,41]. The three classifiers used are logistic regression (LR) [42], support vector machine (SVM) [43], and lightGBM [16].

3.2.4.1. Logistic regression (LR). LR is a statistical model that belongs to linear regression. It allows describing a binomial variable in terms of a collection of random variables, whether categorical or numerical. With additional knowledge about variable values that can be explained or related to that event, it is used to predict the probabilities. Several predicted variables, which might be categorical or numerical, are used in LR. The Logit model or the generic classifier of Entropy is another name for LR. This modeling technique is commonly employed in various scientific and commercial applications, and it is one of the most commonly used modeling techniques in the field of machine learning [42]. LR can be calculated based on Eq. (1) [44]. Figure 4 depicts the separation of data by the decision boundary using the logistic regression classifier.

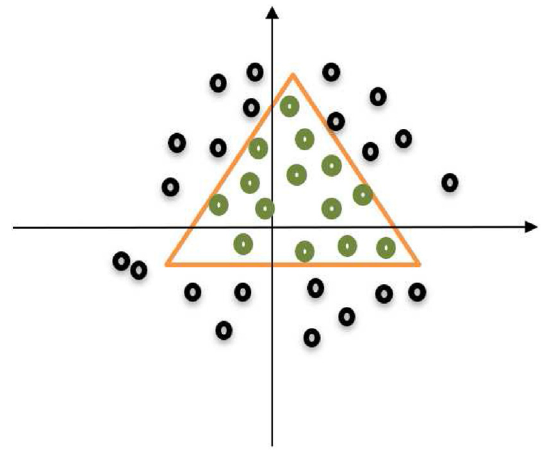

Fig. 4. The separation of data by decision boundary using LR.

$$p = \frac{e^{a+bx}}{1 + e^{a+bx}} \tag{1}$$

where:

p: the predicted value;

e: the base of the natural logarithm (about 2.72);

x: the input value;

a: the bias or intercept term;

b: the coefficient for input (x);

3.2.4.2. Support vector machine (SVM). SVM is a supervised machine learning that can be used to solve issues such as regression and classification. SVM is a good classification approach that can tackle linear and non-linear problems. It is also a practical learning method based on statistical learning theory, so statistical learning theory is the basis of SVM [43]. It can be calculated based on Eq. (2) [45]. Figure 5 shows the separation of data by functional margin based on the support vector classifier.

$$\mathbf{W} = \sum_{i=1}^{N} \alpha_i \mathbf{Y}_i \mathbf{X}_i \tag{2}$$

where:

N: the number of support vectors;

Xi: the input vector of data points closest to the hyperplane;

αi: a Lagrange multiplier of Xi;

Yi: class label;

3.2.4.3. LightGBM. LightGBM is a powerful machine learning technique. It is a gradient boosting framework that is based on a decision tree (DT) which can

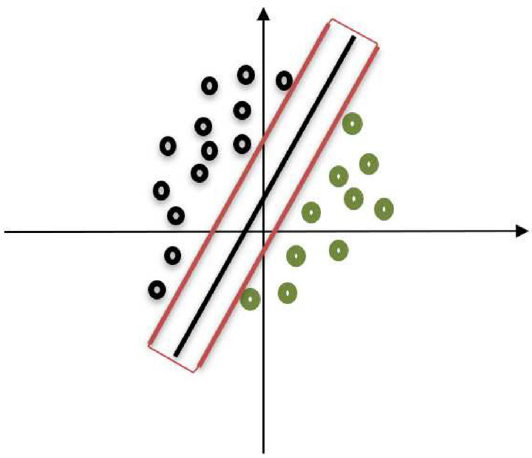

Fig. 5. The separation of data by functional margin using SVM.

provide quick training and high efficiency [46]. LightGBM includes several parameters, termed hyperparameters. The hyperparameters have a significant impact on the performance of the lightGBM algorithm. They are typically set manually. The lightGBM algorithm also helps obtain low memory utilization, large-scale data handling, and high accuracy. It grows vertically, whereas the other algorithms grow horizontally. Although the leaf-wise approach is superior in the level base in terms of minimizing loss and improving accuracy, it is complex and may result in overfitting [16]. Therefore, this technique has many advantages such as quick training, minimal memory usage, high model precision, supporting parallel learning, and being adequate for big data [16].

3.2.4.4. Grid search optimizer. The grid search optimizer method is widely used for determining the appropriate hyperparameters of a classification model. It can possibly reach the ideal solution if there are enough grid nodes [47]. This research integrates the grid search-based optimization into the three algorithms used to select the optimal hyperparameters. By using cross-validation, the dataset is randomly split into test and training sets in the grid search optimizer method [47]. Finding the best hyperparameters can significantly affect prediction accuracy.

3.2.5. Evaluation

Performance measures are calculated to compare the prediction accuracy. Evaluation is the stage of determining if the model is performing well, according to every possible evaluation method, as the accuracy metric is the most commonly used metric for model evaluation in classification [39].

A confusion matrix is a table that can be created for a classifier on a binary dataset and it is used to describe the classifier's performance. The confusion matrix includes four measures which are false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP) (see Fig. 6). In this matrix:

(FN) − The actual is no and the prediction is no.
(FP) − The actual is no and the prediction is yes.
(TN) − The actual is yes and the prediction is no.
(TP) − The actual and the prediction are yes.

4. Results

The accuracy of the prediction performance of the three classifiers was evaluated. The accuracy measure is more sensitive to the distribution of the target variable, as well as the performance of the classifier on an unbalanced dataset. After the preprocessing steps, feature extraction, classification, and performance evaluation stages for all three classifiers were performed. The accuracy obtained for each dimension is summarized in Table 2. Figure 7 shows the prediction accuracy of the MBTI four dimensions based on the three classifiers namely, LR, SVM, and lightGBM. It indicates that the lightGBM classifier performs significantly better than the other classifiers in terms of accuracy. Figure 8 and Fig. 9 illustrate the confusion matrix of lightGBM without and with the implementation of the Grid search optimizer, respectively. The use of the optimizer helps improve the prediction accuracy. However, lightGBM takes a longer time to be executed than other techniques (see Table 3).

The preprocessing stage was used to clean data. It helps remove unimportant words and symbols, return a word to its root form, obtain word stems, remove the sixteen types of personality, omit words that include two letters or less, and delete tags from comments and posts.

The grid search optimizer was also applied which helped select the best hyperparameters to improve the prediction accuracy. The cross-validation also led to balancing the dataset. Such operations

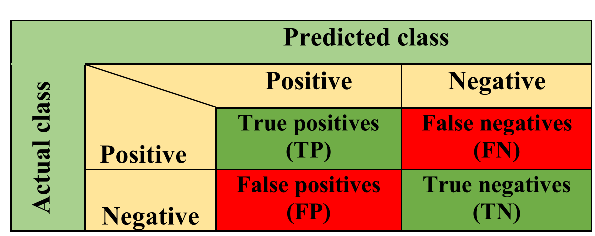

Fig. 6. The confusion matrix.

Table 2. The accuracy after different preprocessing techniques and integrating the Grid search optimizer.

Kaggle dataset from Personality Cafe Forum in 2017	Accuracy											
	Logistic Regression				Support Vector classifier				LightGBM			
	IE	NS	FT	JP	IE	NS	FT	JP	IE	NS	FT	JP
WO stop-words, W stemming, W lemmatizing.	85.01	89.28	85.88	80.29	86.46	90.49	85.48	80.58	86.97	91.82	84.09	82.59
Average of the four dimensions	85.11				85.75				86.36			
W stop-words, W stemming, W lemmatizing.	85.19	90.03	85.82	80.63	86.05	90.72	85.53	80.98	86.92	91.87	84.55	82.94
Average of the four dimensions	85.41				85.82				86.57			
W stop-words, WO stemming, WO lemmatizing.	85.59	89.39	84.55	79.48	86.11	89.91	83.98	79.88	86.05	91.18	83.98	80.81
Average of the four dimensions	84.75				84.97				85.50			
W stop-words, WO stemming, W lemmatizing.	85.24	89.22	85.24	80.06	85.88	90.20	84.50	80.00	86.40	91.01	84.67	81.79
Average of the four dimensions	84.94				85.14				85.96			
W stop-words, W stemming, WO lemmatizing.	85.13	89.97	85.53	80.52	85.99	90.72	85.48	81.04	86.74	91.70	84.50	82.77
Average of the four dimensions	85.28				85.80				86.42			
WO stop-words, WO stemming, WO lemmatizing.	83.75	87.32	83.69	79.08	85.59	88.65	83.80	79.83	85.59	91.07	83.63	81.44
Average of the four dimensions	83.46				84.46				85.43			
WO stop-words, W stemming, WO lemmatizing.	85.07	89.34	85.71	80.23	86.40	90.49	85.19	80.46	86.40	91.24	84.67	82.48
Average of the four dimensions	85.08				85.63				86.19			
WO stop-words, WO stemming, W lemmatizing.	84.78	88.88	84.61	79.88	85.53	89.68	84.15	79.83	85.24	90.55	83.69	80.81
Average of the four dimensions	84.53				84.79				85.07			
W stop-words, W stemming, W lemmatizing, W Grid Search Optimizer (Fit + Score).	87.89	91.35	88.29	86.16	94.52	96.25	94.81	94.69	100.0	100.0	100.0	100.0
Average of the four dimensions	88.42				95.06				100.0			

Note: W=With, WO=Without.

contributed to a significant improvement in forecasting accuracy. Table 2 displays all findings of the algorithms with and without different steps of the preprocessing phase, whereas Table 3 shows the execution time in seconds of all classifiers.

5. Discussion

This research aimed at predicting users' personality types on SMSs by comparing the performance accuracy of three well-known classifiers. It also attempted to improve the prediction accuracy by generating many datasets from the original data using different preprocessing steps. Finally, the research integrated the grid search optimizer with the three classifiers to further enhance the prediction accuracy.

Table 4 compares the performance accuracy of this research with earlier literature. Although previous studies implemented many classifiers, Table 4 includes the findings of a classifier with the highest performance accuracy. It is clear that Naïve Bayes, Gradient Boosting, and SVM achieved the highest accuracy in earlier research. However, the overall findings of this study outperformed the findings of such literature on the same dataset as shown in Table 2. The overall accuracy obtained in this research was 100% based on lightGBM. The rationale behind this may be the implementation of different preprocessing steps before building the classification model. Moreover, the use of the grid search optimizer improved the overall accuracy from about 86% to 100%.

The findings suggest that lightGBM with the implementation of stemming, lemmatization, and removing stop-words as well as integrating the grid search optimizer can produce the best accuracy. Many reasons can be drawn behind such results. First, stemming helps reduce the derived words to their word stem [39], so it decreases the number of words in the corpus and correlates with the word

with similar meanings. Second, another approach used which is similar to stemming is lemmatization, but it compares the words with a language dictionary [39], so this helps the classifiers better predict the labels. Third, the implementation of stop-words removal leads to eliminating low-information terms from the text [40] and this, in turn, reduced the size of the dataset. Finally, the integration of the grid

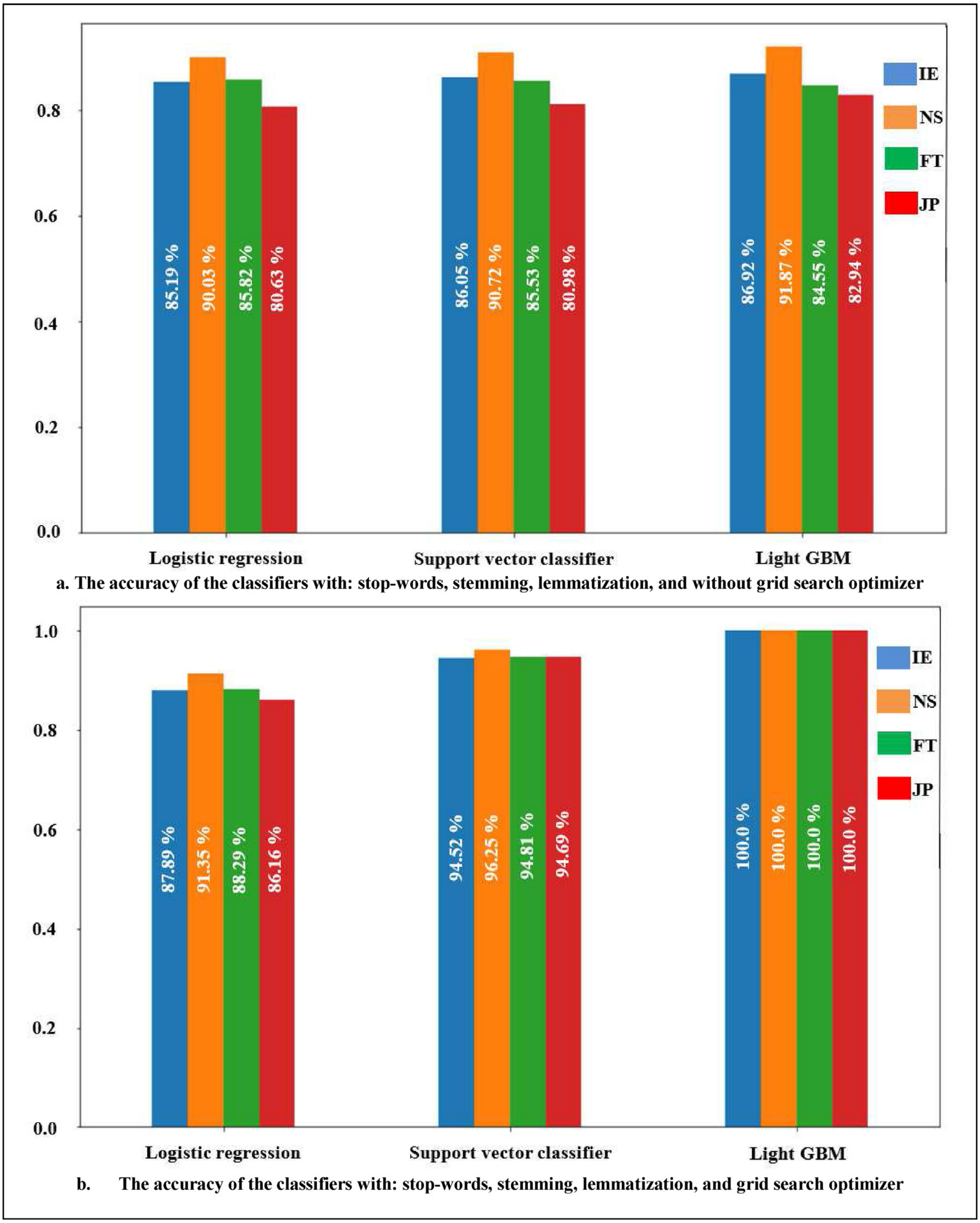

Fig. 7. The accuracy of the classifiers with: stop-words, stemming, lemmatization, and with and without the grid search optimizer.

search optimization identified the best hyper-parameters [47]. Moreover, the K-fold cross-validation was used with the grid search optimizer which splits the data into k parts and ensures that each part is used as a test set to obtain rid of imbalanced data and reduce overfitting.

Following such rigorous procedures in this research resulted in high prediction accuracy. The overall model performance is 100% for the four dimensions, whereas the classifier achieved 86.92%, 91.87%, 84.55%, and 82.94% for (IE), (NS), (FT), and (JP) respectively without the grid search optimization.

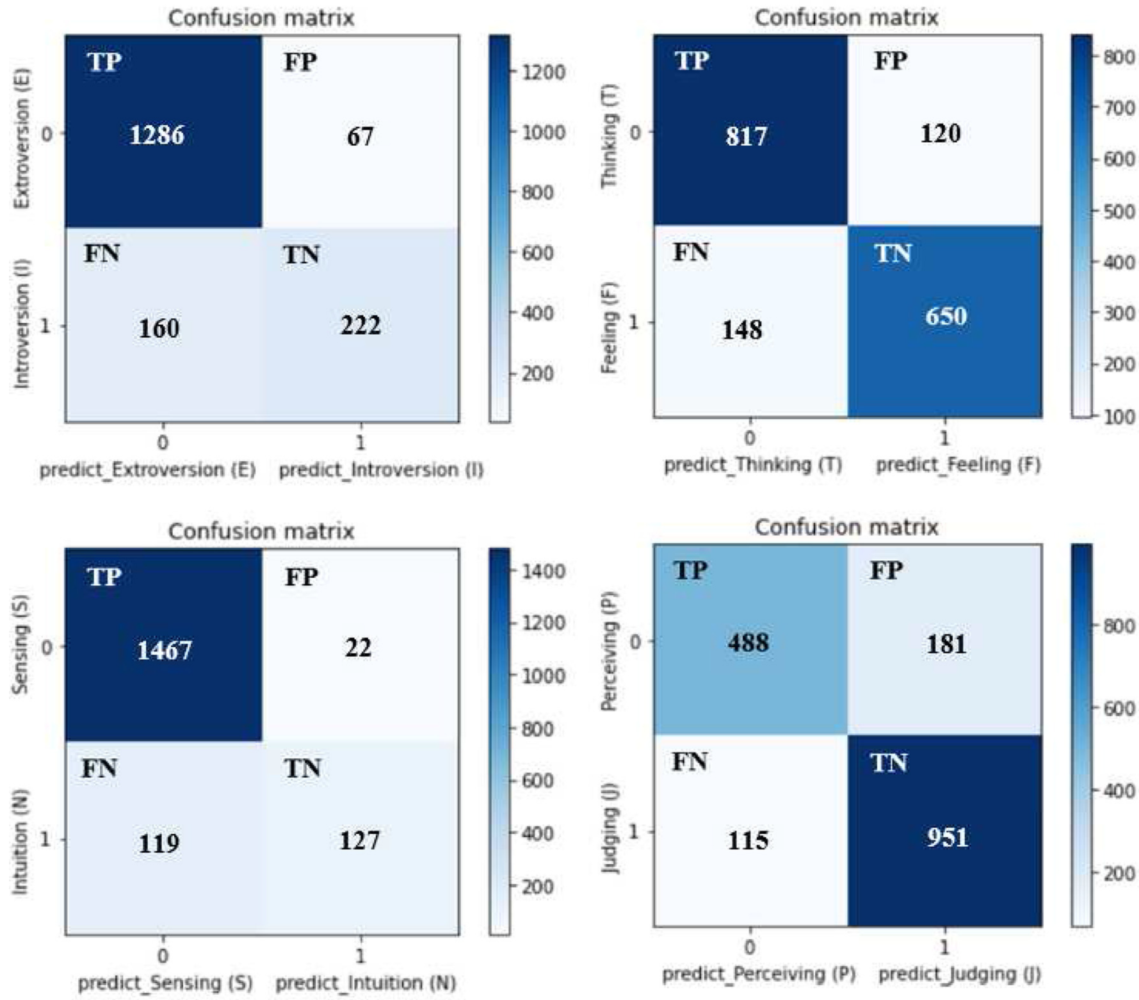

Fig. 8. The confusion matrix with: stop-words, stemming, lemmatization, and without grid search optimizer for the lightGBM algorithm.

Although previous literature [18,25,26] used the same feature extraction methods, they followed different preprocessing steps and data split methods. In another research [16], the lightGBM also showed higher accuracy performance than other techniques, but the grid search optimizer was not used, resulting in lower prediction accuracy than in this study.

Pertaining to the execution time, it was found in our study that lightGBM needs a longer time in comparison with the other classifiers. On the other hand, logistic regression took less execution time than other classifiers, but it showed lower prediction accuracy. Without integrating the grid search optimizer, SVM required a longer execution time, but its accuracy was less than the lightGBM algorithm. After integrating the optimization technique, lightGBM took a longer time than other classifiers. This is because its key concept depends on the procedure of the decision tree as well as selecting a large number of parameters can increase the possibilities of choosing the best parameters and this, in turn, leads to maximizing the execution time.

6. Theoretical implications

The theoretical implications of this research are twofold. First, it follows a rigorous preprocessing technique that helps significantly improve the prediction accuracy of the proposed model. Accordingly, the overall findings of this research outperform related literature. Second, to investigate the effect of optimization on techniques on the prediction accuracy, the study applied the Grid Search optimizer which shows a significant enhancement in the performance accuracy of the four dimensions. Then, the overall results were compared with earlier research based on implementing and not implementing the optimizer.

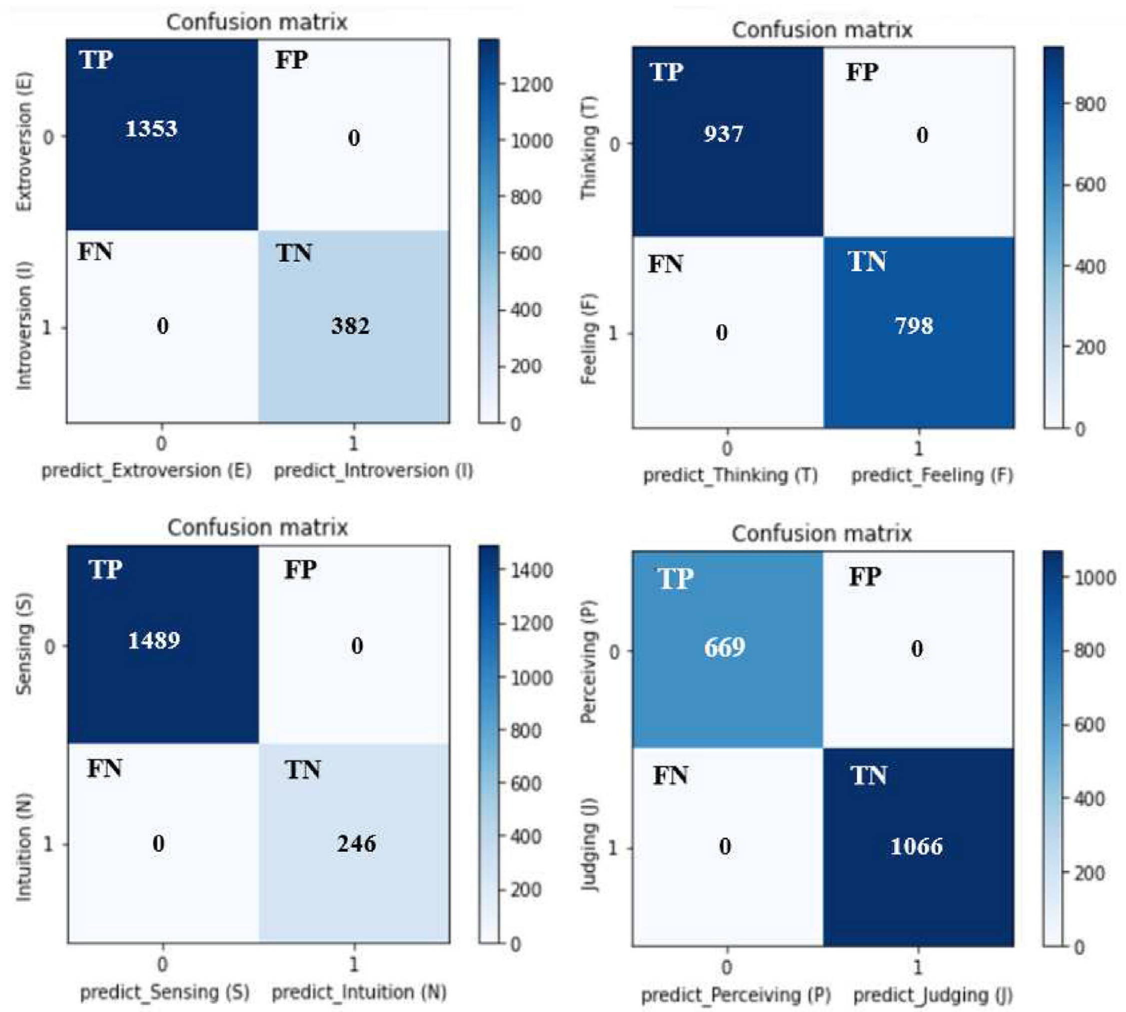

Fig. 9. The confusion matrix with: stop-words, stemming, lemmatization, and grid search optimizer for the lightGBM algorithm.

Table 3. The execution time of all cases.

Kaggle dataset from Personality Cafe Forum in 2017	Logistic Regression	Support Vector classifier	LightGBM
	Execution time	Execution time	Execution time
WO stop-words, W stemming, W lemmatizing.	2.5210 Seconds	140.3919 Seconds	31.0166 Seconds
W stop-words, W stemming, W lemmatizing.	2.4963 Seconds	146.0681 Seconds	30.35016 Seconds
W stop-words, WO stemming, WO lemmatizing.	2.5544 Seconds	153.4194 Seconds	30.5234 Seconds
W stop-words, WO stemming, W lemmatizing.	2.3739 Seconds	150.5366 Seconds	29.7117 Seconds
W stop-words, W stemming, WO lemmatizing.	2.4486 Seconds	147.0819 Seconds	30.06909 Seconds
WO stop-words, WO stemming, WO lemmatizing.	3.6067 Seconds	139.1776 Seconds	36.4450 Seconds
WO stop-words, W stemming, WO lemmatizing.	2.38371 Seconds	140.1468 Seconds	31.3728 Seconds
WO stop-words, WO stemming, W lemmatizing.	2.5342 Seconds	148.5092 Seconds	31.3742 Seconds
W stop-words, W stemming, W lemmatizing, W Grid Search Optimizer (Fit + Score)	59.26813 Seconds	2374.91606 Seconds	19939.0398 Seconds

Note: W=With, WO=Without, S= Seconds.

Table 4. A comparison between the findings of previous research and this study.

Reference	Method	Results			
		IE	FT	SN	JP
[25]	Naïve Bayes	77.0	86.2	77.9	62.3
[26]	Logistic Regression	84.9	88.4	87.0	78.8
[17]	Naïve Bayes	59.0	57.0	63.0	62.0
[18]	XGBoost	78.1	86.0	71.78	65.7
This study without grid search optimizer	LightGBM	86.92	91.87	84.55	82.94
This study with grid search optimizer	LightGBM	100.0	100.0	100.0	100.0

7. Practical implications

This research attempted to improve the accuracy of predicting personality on social media. Identifying people's behavior and personalities has many positive implications on SMSs. First, it can be applied in many sectors that provide different services for individuals such as marketing or educational services. Second, information systems or SMSs can be personalized based on user's behavior. For example, educational hypermedia systems can be adapted based on student's individual needs and this, in turn, can develop their individual motivation and reliability. Third, the research outcomes could also assist organizations in recruiting and selecting the appropriate personality methods that can improve their business by taking into account the personality and preferences of their customers. Finally, identifying users' personalities can lead to choosing the most suitable areas of work that most fit their personalities.

8. Conclusions

This study aimed at implementing machine learning methods to automate personality type prediction based on one of the most used personality models which is MBTI. Natural language processing was used to achieve this aim. The accuracy, time, and performance of the three algorithms were evaluated.

In order to achieve better accuracy and reliability, the research presented a methodology that greatly improved the accuracy of predicting the four personality dimensions of the MBTI model. The accuracy obtained was 100.0% based on the lightGBM Algorithm for the four dimensions. This can actively assist NLP practitioners and psychologists in the identification of personality types and associated cognitive processes on SMSs.

The results of this research proved that personality analysis through text on SMSs is an important factor in predicting the users' personalities. Based on these results, a recommendation system can be built to predict users' personalities on SMSs through their profiles, images, and likes. Regardless of such important outcomes, this research is not without limitations. First, the research was based on one dataset, so it is necessary to apply the proposed model to other SMSs. This can confirm the validity of the existing results. Second, the proposed model was implemented with traditional machine learning techniques, whereas implementing other methods such as deep learning, Monarch Butterfly Optimization (MBO) [48], Earthworm Optimization Algorithm (EWA) [49], Elephant Herding Optimization (EHO) [50], and Moth Search (MS) Algorithm [51] may help provide further research directions.

References

[1] C. Chan, M.J. Holosko, The utilization of social media for youth outreach engagement: a case study, Qual Soc Work 16 (5) (2016) 680−697, https://doi.org/10.1177/1473325016638917.

[2] S. Weshah, E. Alazzam, Q. Aldabbas, Z. Obeidat, M. Humeedat, Y. AlQudah, The impact of social media applications on predicting stock's prices and exchange volume: the case of Jordan, Acad Strat Manag J 20 (6) (2022) 1−10.

[3] X. Luo, Efficient English text classification using selected machine learning techniques, Alex Eng J 60 (3) (2021) 3401−3409, https://doi.org/10.1016/J.AEJ.2021.02.009.

[4] B. Caci, M. Cardaci, M.E. Tabacchi, F. Scrima, Personality variables as predictors of Facebook usage, Psychol Rep 114 (2) (2014) 528−539, https://doi.org/10.2466/21.09.PR0.114k23w6.

[5] J. Marie Condie, I. Ayodele, S. Chowdhury, S. Powe, A.M. Cooper, Personalizing Twitter communication: an evaluation of 'rotation-curation' for enhancing social media engagement within higher education, J Market High Educ 28 (2) (2018) 192−209, https://doi.org/10.1080/08841241.2018.1453910.

[6] J.A. Golbeck, Predicting personality from social media text, AIS Trans Repl Res 2 (1) (2016) 1−10, https://doi.org/10.17705/1atrr.00009.

[7] G. Drakopoulos, A. Kanavos, P. Mylonas, P. Pintelas, Extending fuzzy cognitive maps with tensor-based distance metrics, Mathematics 8 (11) (2020) 1−25, https://doi.org/10.3390/math8111898.

[8] M.Y. Mun, S.Y. Hwang, Impact of DISC behavioral styles on job satisfaction and clinical competencies among newly hired nurses, J Korean Acad Nurs Adm 21 (1) (2015) 43−52, https://doi.org/10.11111/JK 0ANA.2015.21.1.43.

[9] N. Ahmad, J. Siddique, Personality assessment using Twitter tweets, Procedia Comput Sci 112 (1) (2017) 1964−1973, https://doi.org/10.1016/j.procs.2017.08.06 7.

[10] K. Akbar, Y. Jin, M. Mahsud, M. Akbar, A. Waheed, R. Amin, Role of big five personality traits in sustainable consumption behavior, in: ACM int. Conf. Proceeding ser., 2020, pp. 222−226.

[11] A. Behaz, M. Djoudi, Adaptation of learning resources based on the MBTI theory of psychological types, Int J Comput Sci Issues (IJCSI) 9 (1) (2012) 135−141.

[12] A. Eryilmaz, Perceived personality traits and types of teachers and their relationship to the subjective well-being and academic achievements of adolescents, Educ Sci Theor Pract 14 (6) (2014) 2049−2063, https://doi.org/10.12738/estp.2014.6.2187.

[13] R.F. Krueger, N.R. Eaton, Personality traits and the classification of mental disorders: toward a more complete integration in DSM-5 and an empirical model of psychopathology, Pers Disord 1 (2) (2010) 97−118, https://doi.org/10.1037/A0018990.

[14] K. Razavipour, Review of the palgrave handbook of applied linguistics research methodology, Palgrave Handb Appl Linguist Res Methodol 5 (1) (2020) 1–5, https://doi.org/10.1186/S40862-020-00095-X.
[15] J.K. Boparai, S. Singh, P. Kathuria, How to design and validate a questionnaire: a guide, Curr Clin Pharmacol 13 (4) (2018) 210–215, https://doi.org/10.2174/1574884713666180807151328.
[16] E.J. Choong, K.D. Varathan, Predicting judging-perceiving of Myers-Briggs Type Indicator (MBTI) in online social forum, PeerJ 9 (1) (2021) 1–23, https://doi.org/10.7717/peerj.11382.
[17] P. Kadambi, Exploring personality and online social engagement: an investigation of MBTI users on Twitter, University of Chicago, Master dissertation. 2022.
[18] M.H. Amirhosseini, H. Kazemian, Machine learning approach to personality type prediction based on the Myers–Briggs type indicator, Multimodal Technol Interact 4 (1) (2020) 1–15, https://doi.org/10.3390/mti 4010009.
[19] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernández, Syntactic N-grams as machine learning features for natural language processing, Expert Syst Appl 41 (3) (2014) 853–860, https://doi.org/10.1016/j.eswa.2013.08.015.
[20] M.L. Kern, P.X. McCarthy, D. Chakrabarty, M.A. Rizoiu, Social media-predicted personality traits and values can help match people to their ideal jobs, Proc Natl Acad Sci USA 116 (52) (2019) 26459–26464, https://doi.org/10.1073/pnas.1917942116.
[21] A.G. Reece, C.M. Danforth, Instagram photos reveal predictive markers of depression, Springer 6 (1) (2017) 6–15.
[22] R. Safa, P. Bayat, L. Moghtader, Automatic detection of depression symptoms in Twitter using multimodal analysis, Springer US 78 (4) (2022) 4709–4744, https://doi.org/10.1007/s11227-021-04040-8.
[23] L.M. Braunstein, J.J. Gross, K.N. Ochsner, Explicit and implicit emotion regulation: a multi-level framework, Soc Cognit Affect Neurosci 12 (10) (2017) 1545–1557, https://doi.org/10.1093/scan/nsx09 6.
[24] S. Chaudhary, R. Singh, S.T. Hasan, Ms I. Kaur, A comparative study of different classifiers for Myers-Brigg personality prediction model, Int Res J Eng Technol 5 (5) (2018) 1410–1413. https://www.irjet.net/archives/V5/i5/IRJETV5I5268.pdf.
[25] B. Srilakshmi, S. Srinidhi, C. Rahul, S. Ramamoorthy, Persona traits identification based on Myers-Briggs Type Indicator(MBTI) - a text classification approach, in: Int Conf Adv Comput Commun Info (ICACCI), 2018, pp. 1076–1082.
[26] M.S. Raje, A. Singh, Personality detection by analysis of Twitter profiles, in: 8th Int. Conf. Soft Comput. Pattern Recognition 614, 2018, pp. 667–675, https://doi.org/10.1007/978-3-319-60618-7_65.
[27] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, J. Sun, Deep learning-based personality recognition from text posts of online social networks, Appl Intell (2018) 4232–4246.
[28] A. Bhoi, S. Pragyan, P. Rakesh, C. Balabantaray, A deep learning-based social media text analysis framework for disaster resource management, Soc Netw Anal Min 10 (1) (2020) 78–94, https://doi.org/10.1007/s13278-020-00692-1.
[29] M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, Deep learning applications and challenges in big data analytics, J Big Data 2 (1) (2015) 1–21, https://doi.org/10.1186/s4 00537-014-0007-7.
[30] H. Jung, J. Ju, M. Jung, J. Kim, Less-forgetful learning for domain expansion in deep neural networks, Proc AAAI Conf Artif Intell 32 (1) (2018) 3358–3365, https://doi.org/10.1609/aaai.v32i1.11769.
[31] Q. Gong, Y. Chen, X. He, Y. Xiao, P. Hui, X. Wang, X. Fu, Cross-site prediction on social influence for cold-start users in online social networks, ACM Trans Web 15 (2) (2021) 1–23, https://doi.org/10.1145/34091 08.
[32] P. De Meo, E. Ferrara, F. Abel, L. Aroyo, G.J. Houben, Analyzing user behavior across social sharing environments, ACM Trans Intell Syst Technol 5 (1) (2013) 35–51, https://doi.org/10.1145/2535526.
[33] R. Bin Tareaf, P. Berger, P. Hennig, C. Meinel, Cross-platform personality exploration system for online social networks: Facebook vs. Twitter, Web Intell 18 (1) (2020) 35–51, https://doi.org/10.3233/WEB-200 427.
[34] N. Alsadhan, D. Skillicorn, Estimating personality from social media posts, in: 2017 IEEE Int. Conf. Data Min. Work, 2017, pp. 350–356.
[35] L. Liu, D. Preoţiuc-Pietro, Z.R. Samani, M.E. Moghaddam, L. Ungar, Analyzing personality through social media profile picture choice, in: Proc. 10th int. Conf. Web Soc. Media, 2021, pp. 211–220.
[36] A. Furnham, J. Crump, The myers-Briggs type indicator (MBTI) and promotion at work, Psychology 6 (12) (2015) 1510–1515, https://doi.org/10.4236/ps ych.2015.612147.
[37] M. Farrell, Leadership reflections: extrovert and introvert leaders, J Libr Adm 57 (4) (2017) 436–443, https://doi.org/10.1080/01930826.2017.1300455.
[38] R. Müller, M. Martinsuo, T. Blomquist, Project portfolio control and portfolio management performance in different contexts, Proj Manag J 39 (3) (2018) 28–42, https://doi.org/10.1002/pmj.2005 3.
[39] A.S. Khan, H. Ahmad, M.Z. Asghar, F.K. Saddozai, A. Arif, H.A. Khalid, Personality classification from online text using machine learning approach, Int J Adv Comput Sci Appl 11 (3) (2020) 460–476, https://doi.org/10.14569/ijacsa.2020.0110358.
[40] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Syst Appl 38 (3) (2011) 2758–2765, https://doi.org/10.1016/j.eswa.2010.08.066.
[41] L. Kotthoff, I.P. Gent, I. Miguel, An evaluation of machine learning in algorithm selection for search problems, AI Commun 25 (3) (2012) 257–270, https://doi.org/10.3233/AIC-2012-0533.
[42] S. Garg, A. Garg, Comparison of machine learning algorithms for content based personality resolution of tweets, Soc Sci Humanit Open 4 (1) (2021) 100178–100188, https://doi.org/10.1016/j.ssaho.2021.100178.
[43] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, S. Homayouni, Support vector machine versus random forest for remote sensing image classification: a meta-Analysis and systematic review, IEEE J Sel Top Appl Earth Obs Rem Sens 13 (1) (2020) 6308–6325, https://doi.org/10.1109/JSTARS.2020.3026724.
[44] N. sarang, A. Mythili, Classification of soil and crop suggestion using machine learning techniques, Int J Eng Res Technol 9 (2) (2020) 671–673, https://doi.org/10.17577/IJERTV9IS020315.
[45] D. Mining, K. Discovery, B. Laboratories, L. Technologies, A tutorial on support vector machines for pattern recognition, Data Min Knowl Discov 167 (1) (1998) 121–167, https://doi.org/10.1023/A:100 9715923555.
[46] D. Zhang, Y. Gong, The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure, IEEE Access 8 (1) (2020) 220990–221003, https://doi.org/10.1109/ACC ESS.2020.3042848.
[47] G. Erdogan Erten, S. Bozkurt Keser, M. Yavuz, Grid search optimised artificial neural network for open stope stability prediction, Int J Min Reclamat Environ 35 (8) (2021) 600–617, https://doi.org/10.1080/17480930.2021.1899404.
[48] G.G. Wang, S. Deb, Z. Cui, Monarch butterfly optimization, neural comput, Appl 31 (7) (2019) 1995–2014, https://doi.org/10.1007/s00521-015-1923-y.
[49] G.G. Wang, S. Deb, L. Dos Santos Coelho, Earthworm optimisation algorithm: a bio-inspired metaheuristic algorithm for global optimisation problems, Int J Bio-Inspired Comput 12 (1) (2018) 1–22, https://doi.org/10.1504/ijbic.2018.093328.
[50] G.G. Wang, S. Deb, L.D.S. Coelho, Elephant herding optimization, in: 3rd int. Symp. Comput. Bus. Intell., 2015, pp. 1–5.
[51] G.G. Wang, Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems, Memetic Comput 10 (1) (2016) 151–164, https://doi.org/10.1007/s12293-016-0212-3.