A Comparative Analysis of Machine Learning Approaches in Personality Prediction Using MBTI

Kulsum Akter Nisha, Umme Kulsum, Saifur Rahman, Md. Farhad Hossain, Partha Chakraborty, and Tanupriya Choudhury

Abstract A vast amount of data is generated every day on the Internet with the advent of web technology and the increasing number of Internet users. Social media websites such as Twitter and Facebook are increasingly gaining popularity because they help users around the globe to connect and share their opinions on different issues. Personality is how people behave, feel, or think on a specific subject. So using sentimental analysis on Twitter posting the personality of an individual can be easily identified. Twitter data sentiment analysis has been used frequently by lots of researchers for personality classification and prediction. The proposed work also relied on Twitter sentiment analysis to determine a person's personality. In this study, not only the Naive Bayes (NB), Support Vector Machine (SVM), and XGBoost classifier were used to predict the personality of Twitter users but also their performance was compared. Finally, an average of 78% accuracy using NB, 80% accuracy using SVM, and 85% accuracy in XGBoost was found which performed better than others.

Keywords Personality · MBTI · NLP · NB · SVM · XGBoost

K. A. Nisha · U. Kulsum · S. Rahman · P. Chakraborty (✉)
Department of Computer Science & Engineering, Comilla University,
Cumilla 3506, Bangladesh
e-mail: partha.chak@cou.ac.bd

Md. F. Hossain
Department of Statistics, Comilla University, Cumilla 3506, Bangladesh

T. Choudhury (✉)
Informatics Department, School of Computer Science, University of Petroleum and Energy
Studies (UPES), Dehradun 248007, Uttarakhand, India
e-mail: tanupriya1986@gmail.com

A. K. Das et al. (eds.), *Computational Intelligence in Pattern Recognition*,
Advances in Intelligent Systems and Computing 1349,
https://doi.org/10.1007/978-981-16-2543-5_2

1 Introduction

Personality classification is the task of detecting a personality by different categories of measurement. It describes a pattern of thought, feeling, and features that forecasts and illustrates an individual's actions and also influences activities of daily life, such as attitudes, desires, motives, and health. Increased usages of social networking platforms, such as Twitter, Facebook, and Instagram, has encouraged the Internet platform to exchange thoughts, feelings, beliefs, and sentiments, representing their behaviors, actions, and personalities. There is a clear correlation between the personality of the user and the actions seen in the form of social media posts such as tweets on social networking platforms. The attitude behind the text is represented by subjective information: positive and negative.

Nowadays, personality recognition has attracted researchers for developing automatic personality recognition systems. Social networking advantage has produced a wide way of researching sentiment among people and predicting personality. For many types of research, the analysis of individual personality is extremely important. Twitter has become one of the most popular social microblogging platforms on the social network, allowing tweets of up to 140 characters to be read and shared by people. Twitter is an "It's what's happening" website that allows individuals to almost in real time track the views of individual users and comments on different events in their lives [1]. Twitter data follows the data stream model. Twitter users prefer to express their views on the subjects of films, sports, actors, items, social events, and especially trending topics. So that's why we used the Twitter data for our analysis.

In this study, a huge amount of tweets were used to analyze the personality of a person. At the same time, a person cannot be introverted and extroverted. So, according to the Myers–Briggs Type Indicator (MBTI) [2], a person's personality was classified into four category classes of personality. They are Introversion (I)/Extroversion (E), Intuition (N)/Sensing (S), Thinking (T)/Feeling (F), Judging (J)/Perceiving (P). The dataset contained tweet data with a maximum of 50 tweets per person. There are positive and negative feelings for each of the individual classes. For modeling, Naive Bayes, Linear SVM, and XGBoost were used. All the models were evaluated using the different evaluation matrices.

The entire manuscript is organized in such a way that the literature review is done in Sect. 2, the methodology is shown in Sect. 3, the experimental result analysis is in Sect. 4, and the final Sect. 5 is the conclusion.

2 Literature Review

Sentiment analysis is the use of natural language processing, computational linguistics, text analysis, and biometric identification for the purpose of defining, quantifying, collecting, and analyzing effective and subjective information. A lot of work has been done in recent years in the field of Sentiment Analysis [3] by a number of researchers. Nowadays, personality recognition has attracted researchers for developing automatic personality recognition systems. Personality recognition is one kind of Sentiment Analysis. Generally, most of the research work done on sentiment analysis is based on Twitter data. The research primarily focuses on data collection, data preprocessing techniques, and various machine learning algorithms such as NB [4], SVM, and Neural Nets with TF-IDF, LIWC, Emulex, and ConceptNet as feature vectors. The SVM has achieved the highest accuracy in all facets of MBTI for all feature vectors. Logistic Regression, SVM [5], and Multi-Layer Perceptron (MLP) [6] with semantic features of language were used in [7] for predicting personality. Some problems like Demographic factors such as age and gender were not considered.

Pratama and Sarno [8] developed a model to classify user personality in English and Indonesian language using the Big Five Factor Twitter Personality Model. They used KNN, NB, and SVM. The accuracy gained by NB is 60%, KNN is 58%, and SVM is 59%. The model based on NB's accuracy is better than others. One problem is that the final accuracy is not good enough. Using a large dataset and implementing a semantic approach may improve accuracy. The efficiency of the various classifiers was measured in the paper by Chaudhary et al. [9] using the MBTI model to determine a user's personality from the social media texts. For classification and prediction, NB, SVM, Logistic Regression, and Random Forest algorithms were used. The precision obtained by NB, LR [10], and SVM is 55.89%, 66.59%, and 65.44%, respectively. By using a deep learning approach, the outcome can be improved.

Using 1.2 Million tweets, Plank et al. [11] proposed a model for gender and personality prediction. The Twitter dataset was annotated with MBTI personality types. Binary word n-gram and Logistic regression model are used as a feature selection. They used four dimensions of MBTI. Accuracy achieved in personality prediction: I/E = 72.5%, S/N = 77.5%, T/F = 61.2%, and J/P = 55.4%. This analysis indicated an increase in the prediction of I-E and T-F personality groups, but no increase in S-N and a dramatic decrease in J-P. There are a lot of gaps between general personality types and personality types described in the study. The outcome could be improved by adding an optimized dataset. Golbeck et al. [12] developed a model by using the profiles of a Twitter user to reliably identify their personality characteristics using the Big Five Model. A total of 50 person tweets were used for classification and prediction, with 2,000 tweets per person. Two regression models were used, ZeroR and GP. Accuracy was higher for Open with 75.5% and lower for Neuro with 42.8%. A system [13] was encoded for 12 different language characteristics and has established a connection between user attitude and writing style for different users and devices across regions. Compared to the iPhone, Blackberry, Uber social users, and Facebook users, Twitter users have been recognized to be stable, neutral, and introverted. This is an Unsupervised Score-based model. More Twitter classified data can improve the efficiency of the personality identification model.

Celli and Rossi [14] proposed a method to illustrate how different personalities interact and operate on Twitter's social networking platform. Statistical and Linguistic characteristics are used in this work and then tested by human judgment on data corpus elaborated with a personality model and achieved 78.29% Accuracy. An unsupervised method was proposed by Celli [15], based on the Big Five Personality Model for personality prediction. Different social media network data were used for the extraction and assessment of an individual's personality. An extended annotated corpus can boost the system's performance. Dejan Markovikj et al. [16] tried to find the personality of an individual by extracting data from Facebook. They used SVM to find a result. In exploratory scenario-based case studies for chosen domains, the long-term goal is to prove the effectiveness of predictive models.

3 Methodology

First, a publicly accessible standard dataset [17] was collected from Kaggle. Then, the dataset was preprocessed by tokenizing, removing stopwords, and vectorization. We have designed three supervised models, namely Naive Bayes, SVM, and XGBoost in our study. After that, we trained the model with the training dataset. Finally, the proposed model was evaluated by creating evaluation matrices such as confusion matrices, F1-score, and accuracy. The sequential steps of the full methodology are shown in Fig. 1.

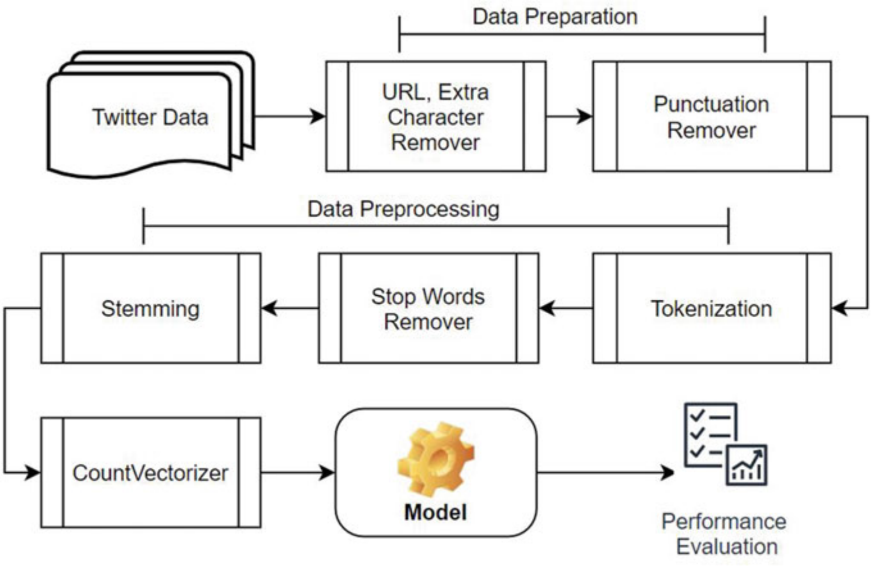


Fig. 1 Proposed methodology

3.1 Dataset Description

A publicly accessible standard dataset [17] was collected for the proposed study. The dataset was made by scrapping people's tweets. In the dataset, each user has at most 50 tweets, and it consists of 8675 rows where each row represents a specific person. The Personality of each user was annotated by the Myers–Briggs Type Indicator (MBTI). The MBTI system divides personalities into four different groups and 16 different types. The initial dataset is completely skewed and unevenly distributed among all four groups. The following four groups are shown.

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

Though this dataset contains 16 types of personalities, they are not equal in number. Here, INFP has the most frequency with 1832 rows, and ESTJ has the lowest with 39 rows shown in Fig. 2.

3.2 Data Preparation

A tweet generally includes opinions about something or someone that is expressed by various users in various forms. Raw data is often unformatted and unstructured and can cause a decrease in the model's performance. So, the data was formatted and cleaned. To get an accurate result for in the model's performance, all unnecessary words and punctuation was processed. In the case of cleaning the data, these steps were followed.

- All URLs (e.g. www.abc.com) were removed.

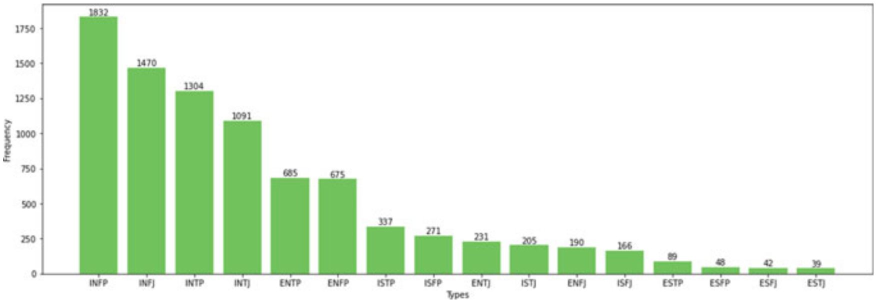


Fig. 2 Frequency of MBTI types

Table 1 Statistical information of dataset

Total words count	10870272
Total characters count	56791898
Maximum words in a row	1923
Maximum characters in a row	9557
Top 5 words	the, to, and, of, you
Frequency of top word	308578
Unique vocabulary	146297

- All the emoticons were removed.
- All the punctuation marks, symbols, and numbers were removed.
- All non-English character were removed.
- Target names (@someone) and hashtags were removed.

Some statistical information such as Total words count, Total characters count, and Top frequent words of the cleaned dataset are shown in Table 1.

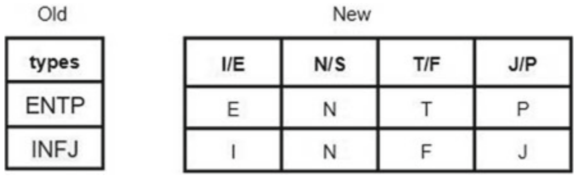
3.3 Data Preprocessing

The Twitter dataset was preprocessed in three steps before fitting into the classifiers. Preprocessing is necessary for text processing because it transforms the data into a standard form before turning them into features. It's like making the text into a machine-readable format from a human-readable format.

Tokenization The raw text is divided into small blocks by Tokenization. Tokenization splits the raw text into words or sentences called tokens. Such tokens assist in interpreting the meaning of the text for creating an NLP model. There are generally two types of tokenization. The Word Tokenizer splits the text into words and the Sentence Tokenizer splits the text into sentences. We used a word tokenizer in the proposed study.

Stop Words Removing There are some words in every natural language that are most common but do not add much meaning to the sentences. Those words like the, to, and, etc. are called stop words. Removing stop words can cause faster training time and improved model performance. Stop words were removed from the dataset for their less impactive value.

Stemming New words can be formed by adding suffixes, prefixes, or inflections to the root word. Stemming is the process of reducing a word to its base form by removing these suffixes, prefixes, or inflections. For example, 'runs', 'ran', 'running' all of them will be reduced to their base form 'run'. Stemming can reduce the complexity of the dataset by decreasing the size of the vocabulary thus giving faster training time.



Old

types
ENTP
INFJ

New

I/E	N/S	T/F	J/P
E	N	T	P
I	N	F	J

Fig. 3 Personality group splitting

For four different personality groups, we split each group of the personality into a distinctive column so that it will be easy to predict one group at a time, which is shown in Fig. 3.

3.4 Feature Extraction

CountVectorizer is a feature extraction method in text processing. It takes a tokenized text dataset as inputs. CountVectorizer makes a sparse matrix by making a vocabulary including all unique tokens and counting their occurrences.

3.5 Classification Model

After preprocessing and feature extraction, three supervised classifiers named NB, SVM, and XGBoost were designed for training. The classification algorithms that were used in this study are detailed below.

Naive Bayes (NB) NB is the most fastest classifying algorithm that uses the Bayes probability theorem to predict an unknown class. Bayes' Theorem also determines the probability of an event occurring, given the probability of another event that has already occurred. The classifier of Naive Bayes assumes that the effect of a particular feature in the class is independent of other features. Even if these features are interdependent, they are still considered independent.

$$P(C|d) = \frac{P(d|C)P(C)}{P(d)} \tag{1}$$

In Eq. 1, C is the personality class, d is features as tweets, and $P(C|d)$ is the probability of a personality class given feature vector d.

Support Vector Machine (SVM) SVM is a supervised learning technique that analyzes data used for regression and classification analysis using a series of kernel-defined mathematical functions, and the kernel's job is to take data as input and transform it into the appropriate form. The SVM approach works by representing different classes in a hyperplane of multidimensional space. In the proposed study,

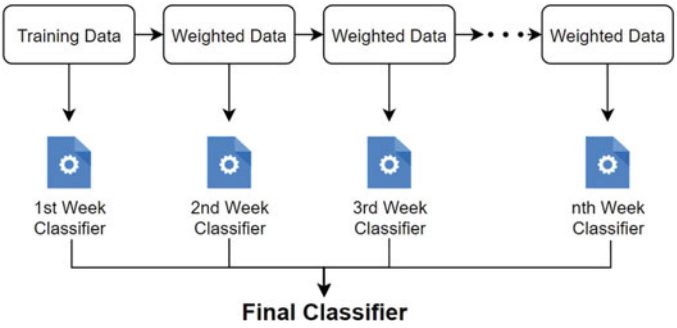


Fig. 4 XGBoost classifier

the Linear kernel of SVC was used. The formula of the linear kernel is shown in Eq. 2,

$$S(d, d_i) = sum(d * d_i) \tag{2}$$

XGBoost XGBoost is a decision-tree-based machine learning ensemble algorithm that uses a gradient-boosting process optimized for speed and performance. In XGBoost, weights play a significant part. All the independent variables are then fed into the decision tree that forecasts outcomes with given weights. The weight of variables that are estimated incorrectly by the tree is adjusted and then fed into the second decision tree. These individual classifiers/predictors then tie together to form a more efficient and comprehensive model, as shown in Fig. 4. Regression, grouping, rating, and user-defined prediction issues can be easily solved by the XGBoost algorithm.

The dataset was split into two parts, train dataset and test dataset. The testing data was 25% of the dataset which contained 2169 rows of tweets. In Naïve Bayes, multinomial NB kernel, and in SVM, Linear SVC kernel was used. For the XGBoost classifier, the ‘objective’ was set to ‘binary:logistic’ as for each group of personality there were only two classes. Each of the classifiers was trained four times for the four groups of personality. After training each classifier with the training dataset, they were evaluated using the testing dataset. To get the best performance, some of the parameters of the classifiers were tuned. The best set of parameters used in the proposed model are given below.

- NB: we used default set of parameters for NB classifier.
- SVM: optimal set of parameters is ‘C’: 10; ‘gamma’: 0.1; ‘kernel’: ‘linear’.
- XGBoost: optimal set of parameters is ‘objective’: ‘binary:logistic’; ‘max depth’: 5; ‘learning_rate’: 0.1; ‘alpha’: 10.

4 Experimental Result Analysis

All the classifiers' performance was evaluated using four evaluation measurements such as precision, recall, F1-score, and accuracy. These evaluation measurements were calculated from the confusion matrix. The confusion matrix for XGBoost Classifier is given in Table 2. The same thing was done for the other two classifiers also.

Precision, Recall, and F1-score measure for all three classifiers are shown in Table 3. Here, the XGBoost classifier gave better results than others. Its average F1-score measure is higher and achieved the highest F1-score of 0.88 in predicting the N/S personality group.

Here, XGBoost classifier achieved the highest with 85% average accuracy which is given in Table 4. Using NB, accuracy attained for I/E = 76%, N/S = 85%, T/F = 80%, and J/P = 73%.

Linear SVC classifier achieved I/E = 80%, N/S = 86%, T/F = 80%, and J/P = 72% accuracy. Using XGBoost classifier, accuracy achieved for I/E = 86%, N/S = 90%, T/F = 84%, and J/P = 80%. An average accuracy of 78% was achieved using Naive Bayes and an average accuracy of 80% was achieved using Support Vector Machine from all the four personality groups.

Table 2 Confusion matrix for XGBoost Classifier

Groups	Actual → Predicted ↓	I	E
I/E	I	1578	65
	E	256	270
N/S		N	S
	N	1840	34
	S	191	104
T/F		T	F
	T	854	171
	F	187	957
J/P		J	P
	J	550	300
	P	145	1174

Table 3 Result analysis

PG	NB			SVM			XGBoost		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
I/E	0.74	0.78	0.73	0.80	0.81	0.81	0.85	0.85	0.85
N/S	0.78	0.85	0.80	0.88	0.88	0.88	0.89	0.90	0.88
T/F	0.81	0.81	0.81	0.79	0.79	0.79	0.84	0.83	0.84
J/P	0.75	0.75	0.75	0.75	0.74	0.74	0.79	0.79	0.79

*Here, PG means Personality Groups

Table 4 Accuracy of classifiers for each personality class

Personality group	Accuracy		
	NB (%)	SVM (%)	XGBoost (%)
I/E	76	80	86
N/S	85	86	90
T/F	80	80	84
J/P	73	72	80

5 Conclusion

The proposed system achieved satisfactory results in predicting the personality by using tweets. In three classifiers, the Naive Bayes classifier's performance is worse while Linear SVM and XGBoost give a satisfactory result. These results were with an unbalanced dataset. So the performance of these classifiers can be improved by making the dataset balanced or creating a new balance dataset. Ideally, Gender can be a key factor in the differentiation of the form of personality based on the choice of word. Age, education level, nationality, first language, country, and even religion can also have an effect on word choice that could help to improve the performance of any classifier. Considering this issue, the proposed model can be optimized for future implementation. Using more data from other social media sources like Facebook, Instagram, etc. can add a better perspective and performance in predicting personality.

References

1. Schonfeld, E.: Mining the Thought Stream. Tech Crunch Weblog Article (2009). http://techcrunch.com/2009/02/15/mining-the-thought-stream/
2. Bharadwaj, S., Sridhar, S., Choudhary, R., Srinath, R.: Persona traits identification based on myers-briggs type indicator(MBTI)—a text classification approach. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Bangalore, pp. 1076–1082 (2018)
3. Rahman, S., Chakraborty, P.: Bangla document classification using deep recurrent neural network with BiLSTM. In: Proceedings of International Conference on Machine Intelligence and Data Science Applications (2020). https://doi.org/10.1007/978-981-33-4087-9_43
4. Ahammad, K., Chakraborty, P., Akter, E., Fomey, U.H., Rahman, S.: A comparative study of different machine learning techniques to predict the result of an individual student using previous performances. Int. J. Comput. Sci. Inf. Secur. **19**(1), 5–10
5. Zulfiker, M.S., Kabir, N., Biswas, A.A., Chakraborty, P., Rahman, M.M.: Predicting students' performance of the private universities of Bangladesh using machine learning approaches. Int. J. Adv. Comput. Sci. Appl. **11**(3), 672–679 (2020)
6. Chakraborty, P., Yousuf, M.A., Rahman, S.: Predicting level of visual focus of human's attention using machine learning approaches. In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering (2021). https://doi.org/10.1007/978-981-33-4673-4_56

7. Gjurković, M., Šnajder, J.: Reddit: a gold mine for personality prediction. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions. Personality, and Emotions in Social Media (2018)
8. Pratama, B.Y., Sarno, R.: Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In: 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE (2015)
9. Chaudhary, S., Sing, R., Hasan, S.T., Kaur, I.: A comparative study of different classifiers for myers-brigg personality prediction model. IRJET **05**, 1410–1413 (2018)
10. Chakraborty, P., Zahidur, Md, Rahman, S.: Movie success prediction using historical and current data mining. Int. J. Comput. Appl. **178**(47), 1–5 (2019)
11. Plank, B., Hovy, D.: Personality traits on twitter–or–how to get 1,500 personality tests in a week. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity. Sentiment and Social Media Analysis, pp. 92–98 (2015)
12. Golbeck, J., et al.: Predicting personality from twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE (2011)
13. Celli, F.: Mining User Personality in Twitter. Language, Interaction and Computation CLIC (2011)
14. Celli, F., Rossi, L.: The role of emotional stability in Twitter conversations. In: Proceedings of the Workshop on Semantic Analysis in Social Media, Association for Computational Linguistics, pp. 10–17 (2012)
15. Celli, F.: Unsupervised personality recognition for social network sites. In: Proceedings of Sixth International Conference on Digital Society, pp. 59–62 (2012)
16. Markovikj, D., Gievska, S., Kosinski, M., Stillwell, D.: Mining facebook data for predictive personality modeling. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013). Boston, MA, USA, pp. 23–26 (2013)
17. Jolly, M.: (MBTI) Myers-Briggs Personality Type Dataset. https://www.kaggle.com/datasnaek/mbti-type