

Linguistic feature selection for personality trait identification from textual data

Angad Singh¹, Priti Maheshwary¹, Nitin Kumar Mishra², Timothy Malche³

¹Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, India

²School of Computing Science and Engineering, VIT Bhopal University, Sehore, India

³Department of Computer Applications, Manipal University Jaipur, Jaipur, Rajasthan, India

Article Info

Article history:

Received Jul 8, 2024

Revised Sep 28, 2024

Accepted Oct 7, 2024

Keywords:

Chi-square

Feature selection

F-statistic

Genetic algorithm

Mutual information

Personality trait identification

Principal component analysis

ABSTRACT

Personality identification is a common and central problem in text processing. Sensing personality is helpful for various purposes; for example, estimating users' personalities before providing them with any service is necessary. Individuality is essential in a person's nature in every outlook, for instance, in text writing. But, this remains a core challenge because of the low accuracy achieved. The proposed study solves this problem and presents a big five trait identification technique from text data, which applies a feature selection method to increase accuracy. This technique is called linguistic feature selection for personality trait identification (LFSPTI). This technique first finds features based on mutual information (MI), F-statistic, principal component analysis (PCA), and chi-square, then uses the genetic algorithm (GA) to select high-ranked features from all feature subsets. These four parameters provide various forms of the dataset. The experimental results exhibit that the LFSPTI method enhances the classification accuracy against the best of the competing methods by 1.18%, 0.83%, 1.61%, 1.15%, 1.82%, and 1.39% for extraversion (EXT), neuroticism (NEU), agreeableness (AGR), conscientiousness (CON), openness (OPN), and mean overall personality traits, respectively.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Angad Singh

Department of Computer Science and Engineering, Rabindranath Tagore University

Bhopal, India

Email: angada2007@gmail.com

1. INTRODUCTION

Personality is the combination of all those activities that can be discovered by actual observation over a long period of time to give useful information. Personality describes a person from multiple perspectives, such as thoughts, behavior, beliefs, feelings, and emotions [1]. According to psychology, every personality has a significant influence such that it affects life priorities, well-being, health, and many other choices in daily life. Personality is an enduring set of tendencies and styles that an individual exhibits in his or her life. It reflects individuals' inclinations and differentiates them from the "ideal individual" in the general public or society. Here, the characteristic examines the commitment and personal emotional attitude in a particular type of behavior. There are two commonly used models for personality classification. First, a big-five model consists of five personality traits [2], [3]: extraversion versus introversion, agreeableness versus unpleasantness, conscientiousness versus carelessness, neuroticism versus equanimity, and openness versus closedness. Second, there is the myers-briggs type indicator (MBTI) model [3]-[5], which has four personality traits. Correlation between big-five model and MBTI model shown in [3]. However, the big five trait model

is standard for deep learning and machine learning paradigms because of its proper linguistic personality behavior [6]-[9].

Automation of personality trait recognition has multiple applications, such as in human resource management, decision-making, marketing strategy, tourism, mental health, and forensics. Personality traits help human resource managers to select the right employees for specific positions in human resource management [1]. In decision-making, it helps one to understand the behavior of employees in specific situations. In mental health, contextual analysis is related to certain personality traits. In forensics, a relevant personality trait helps narrow the area of suspicion. In marketing, personality traits influence marketing strategies for consumer goods. Personality trait recognition is imagined as an excellent technique to address the issues and is a promising method in all mentioned applications. Nevertheless, personality trait identification techniques notice a small accuracy score. Various deep learning and machine learning paradigms have recently been suggested to solve this limitation [6].

The big-five trait model is widely used as a personality trait model in personality classification. It is a well-analyzed and observed model of personality architecture that most researchers use. Therefore, most studies use the big-five trait model regarding personality taxonomy. Extract user information to recognize the person's personality based on the behavior of different users. However, personality identity is conventionally found in a user's status updates, such as text messages and profile data. The majority of this information comes from text form. Therefore, the analysis of textual data allows for the judging of essential personality traits [8].

The focus of this work is the prediction of personality traits from textual data. There are various machine learning models that can predict the personality traits of an individual. Feature selection techniques have been applied in some recent works to increase accuracy; however, low accuracy remains a core challenge. This paper addresses the problems by integrating feature extraction from textual data, feature selection, and genetic algorithm (GA).

In recent times, many research studies have been based on machine learning and deep learning approaches for personality recognition [10]. Personality recognition from social networks is widely accepted, with Facebook prominently considering sentiment analysis. Nevertheless, most personality research focuses on feature extraction techniques based on data collection or preprocessing. In traditional machine learning, classification algorithms and multi-layer neural networks determine personality identification. Practically, text features, for example, one-gram, two-gram, or linguistic features, or integration of both linguistic and textual features, are used in all personality models. Instead of selecting the entire set of features, the prediction model performance can be improved using feature selection techniques [11], [12] on smaller datasets.

Past research on personality prediction has been done using the Facebook social platform and a few relevant features such as linguistic inquiry word count (LIWC) features, time-related features, social network analysis (SNA) features, and more [13], [14]. Linguistic Etiquette [15] was a major work demonstrating that composed language is unique for every individual's personality traits, perhaps recognized by using language cues. LIWC [1] is a mechanism of text analysis that counts the frequency of various groups of words in a given text. Investigators have discovered notable correlations between personality traits and linguistic characteristics. Another research on personality prediction [16] was done using Facebook status and features like open vocabulary differential language analysis (DLA) and LIWC [17], [18]. Another research study used Twitter data for personality prediction with features like LIWC and medical research council (MRC) [19].

All the above research studies predicted personality using social media platforms in English using the big-five personality model. Mairesse *et al.* [20] use conversational and text datasets in experiments with LIWC for the big-five personality trait model. The Mairesse criterion relies on the correlation between personality traits and linguistic features for textual datasets. It mined 102 features: 14 MRCPD [21] features from the MRC Psycholinguistic Database and 88 LIWC features from the LIWC mechanism. FineEmotion [22] is an emotion-based hashtag technology. It considers six emotions: happiness, anger, fear, surprise, sadness and disgust. FineEmotion applies word-based sentiment pools to increase personality prediction accuracy on Essay datasets. It is also important to consider that emotions have a powerful relationship with personality traits.

LIWC and feature reduction (FR) is a personality trait recognition technique [23]-[25] using feature reduction and LIWC features. It uses support vector machine (SVM) and LR for classification, and for feature reduction, it uses principal component analysis (PCA) and information gain. It works in such a way that dimensionality reduction reduces time complexity and improves the classifier accuracy. Mishra *et al.* [1] proposed a personality classification method in the form of multi-label classification to establish linguistic behavior and feature selection. It is a model of feature selection that depends on the ranking process. It then selects the best high-ranked features from the merged feature set. The merged feature set is a union of LIWC and MRC features that applies mutual information (MI), chi-square, and F-statistic. It is a label-wise search technique. First, it mines LIWC and MRC features for every label; then, the indicated features are merged to

find a one-feature subset. The high-level features are selected from the merged feature subset. F-statistics, chi-square, and MI analysis of variance were applied to classify features using the wrapper method to obtain better feature sets from diverse features. The feature relationships used in various popular methods are shown in Table 1.

Table 1. Features with respect to various techniques

S.no.	Title	Feature	Classifier used	Results (accuracy)
1.	Mairesse Criterion [20]	102 (LIWC+MRC)	SVM	56.97%
2.	FineEmotion [22]	102 (LIWC+MRC) + hashtag lexicon (585)	SVM	57.64%
3.	LIWC &FR [23]	56 (PCA) and 10 (IG)	LR, SVM	57.92%
4.	PTLFM technique [1]	102(LIWC+MRC)	SVM, LR	60.22%

Theoretically, every one of the given techniques is similar. They use each feature, for example, LIWC only or both LIWC and MRC integrated. As a result, the achieved accuracy is still low and can be improved. To solve this problem, we propose the LFSPTI method, which utilizes LIWC, EMPATH [26], and MRC features. The purpose of this method is to increase accuracy for personality traits.

The major contributions of this research study are mentioned below.

- This work presents a unique approach called linguistic feature selection for personality trait identification (LFSPTI) for mining features.
- It extracts features like LIWC, MRC, and EMPATH from the textual datasets.
- It applies feature selection methods: F-statistic, chi-square, MI, and PCA.
- It uses the GA to select high-ranking features from all the feature subsets.
- Next, it applies machine learning models, SVM, logistic regression (LR), and decision tree (DT) to the selected feature set.
- The results on the popular Essay dataset show that the LFSPTI method's accuracy outperforms the current personality identification methods.

In text computing, this is unique research work that associates LIWC, EMPATH and MRC features according to our past and present knowledge. Then GA based feature selection techniques including F-statistics, chi-square, MI and PCA are applied to choose high ranked features for trait recognition in personality computing. The remaining part of the paper is organized as follows. Section 2 outlines the related concepts and working of the proposed method along with the details of the dataset used in the experimentation. Section 3 describes the results and experimental discussion. Finally, section 4 concludes with possible future research scope.

2. METHOD

This portion describes the proposed LFSPTI method. The method relies on MRC, EMPATH, LIWC features, F-statistics, chi-square, MI, and PCA. So, the first two sub-sections describe these concepts, and the last sub-section describes the LFSPTI method.

2.1. Features (MRC, EMPATH and LIWC)

This work incorporated many features and investigated the relevant capability and performance of various attributes for personality computing. We used MRC, EMPATH, and LIWC linguistic features for this method. The mined MRC features include the number of letters, syllables, intonation, conciseness, and familiarity ratings. EMPATH is a text analysis tool that draws meaning between words and phrases and generates new lexical categories on demand from a small set of seed words. We have also designed LIWC features like word count, words per sentence, positive emotions, anger, work, job, home, comma, colon, and apostrophe, similar to the Mairesse baseline.

2.2. Feature selection techniques

F-statistic is a statistical test that is applied in hypothesis testing to check whether the variance of two samples data is equal or not. This test uses the F-statistic to compare two variances by dividing them. In this context, to compute F-statistic with respect to the x^{th} feature corresponding to the y^{th} personality trait and, it selects appropriate features from x^{th} feature using (1) and inappropriate features from x^{th} feature using (2).

$$A_x^{1(y)} = \{ I_k : J_k = 1 \text{ for } k = 1, 2, 3, \dots \dots n \} \quad (1)$$

$$A_x^{0(y)} = \{ I_k : J_k = 1 \text{ for } k = 1, 2, 3, \dots \dots n \} \tag{2}$$

Further, we calculate *Mean M1* of the set $A_x^{1(y)}$ and *Mean M0* of the set $A_x^{0(y)}$ using (3) and (4).

$$\text{Mean M1} = \frac{\sum_{I \in A_x^{1(y)}} I}{n1} \text{ where } n1 = |A_x^{1(y)}| \tag{3}$$

And

$$\text{Mean M0} = \frac{\sum_{I \in A_x^{0(y)}} I}{n0} \text{ where } n0 = |A_x^{0(y)}| \tag{4}$$

Now, finally (5) computes x^{th} features F-statistic.

$$S_x^{(y)} = \frac{(M0-M)^2 + (M1-M)^2}{N1+N2} \text{ where } M \text{ is overall mean of } A_x, N1 = \frac{1}{n0} \sum_{I \in A_x^{0(y)}} (I - M0)^2 \text{ and } N2 = \frac{1}{n1} \sum_{I \in A_x^{1(y)}} (I - M1)^2 \tag{5}$$

- a) Chi-square method is a statistical technique used to calculate if two categorical variables are related or independent. To compute chi-square with respect to the x^{th} feature corresponding to the y^{th} personality trait using (6):

$$A_x^{(y)} = \sum_{i \in \{0,1\}} \frac{(A_{o(x,i)}^{(y)} - A_{e(x,i)}^{(y)})^2}{A_{e(x,i)}^{(y)}} \tag{6}$$

where $A_{o(x,i)}^{(y)}$ is observed values of x^{th} feature for the y^{th} personality trait and $A_{e(x,i)}^{(y)}$ is expected values of x^{th} feature for the y^{th} personality trait and i denotes the relevancy of the instances.

- b) MI is a statistic to measure the relationship between two random variables that are sampled simultaneously. It determines the dependence of two variables and is computed using (7):

$$MI(X, Y) = \text{Entropy}(X) - \text{ConditionalEntropy}(X / Y) \tag{7}$$

The disorder available in data is determined by entropy. It is computed by (8) for a random variable X.

$$\text{Entropy}(X) = - \sum_{i=1}^n P(Xi) * \log(P(Xi)) \text{ where } P(Xi) \text{ is the probability} \tag{8}$$

If two random variables rely on one another, the entropy of one variable decreases after noticing the second variable. The resting uncertainty of a variable on another is established by conditional entropy. It is calculated by (9) concerning variable X and variable Y

$$\text{ConditionalEntropy}(X / Y) = - \sum_{j=1}^m \sum_{i=1}^n P(Yj) P(Xi/Yj) * \log\left(P\left(\frac{Xi}{Yj}\right)\right) \tag{9}$$

where $P(Xi/Yj)$ is known as the conditional probability of Xi and Yj is given.

- c) PCA is a very famous statistical method for data ordination and dimensionality reduction of data. It is a machine learning method that simplifies a huge data set into a smaller one. PCA can be broken down into the following steps:

- Step 1: Standardize the limit of continuous initial variables.
- Step 2: Compute the covariance matrix to identify correlations between variables.
- Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.
- Step 4: Determining which components are relevant to the analysis by creating a feature vector.
- Step 5: Rearrange the data along the principal component axes.

2.3. Proposed method

The training process of LFSPTI method for y^{th} personality trait is shown in Figure 1 and Algorithm 1. First, it gets all the features of LIWC (X_L), features of MRC (X_M), and features of EMPATH (X_E) from the

input training dataset TR. Then combine LIWC, MRC, and EMPATH features to find a feature set $S=X_L \cup X_M \cup X_E$. Then, it computes MI, chi-square, F-statistic, and PCA for each feature from feature set S using GA to find best high ranked features for personality classification.

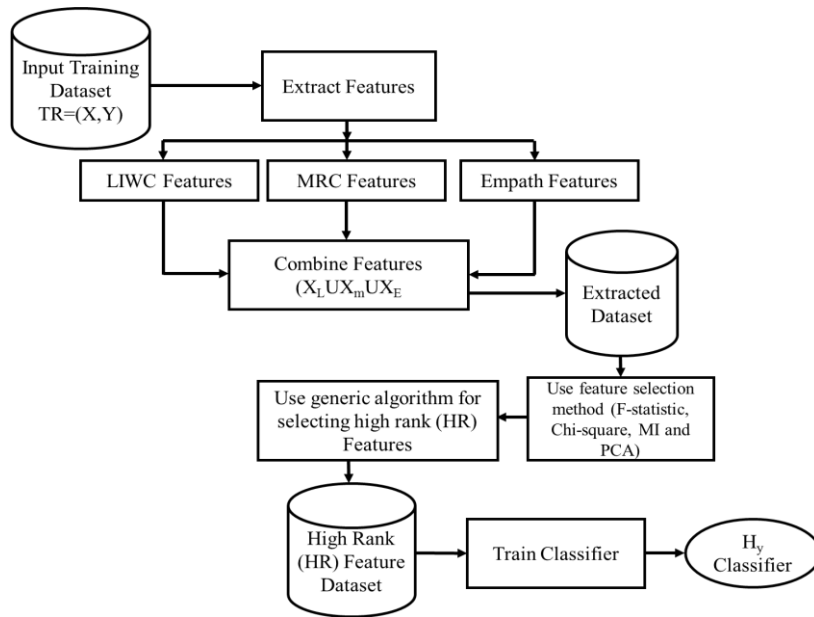


Figure 1. Process of training model

The testing process of LFSPTI method for y^{th} personality trait is shown in Figure 2. Analogous to training, first, it gets all the features of LIWC (X_L), features of MRC (X_M), and features of EMPATH (X_E) from the input test dataset TS. Then, it selects the high-ranked features, as decided during training, to form the extracted dataset for predicting the y^{th} personality trait.

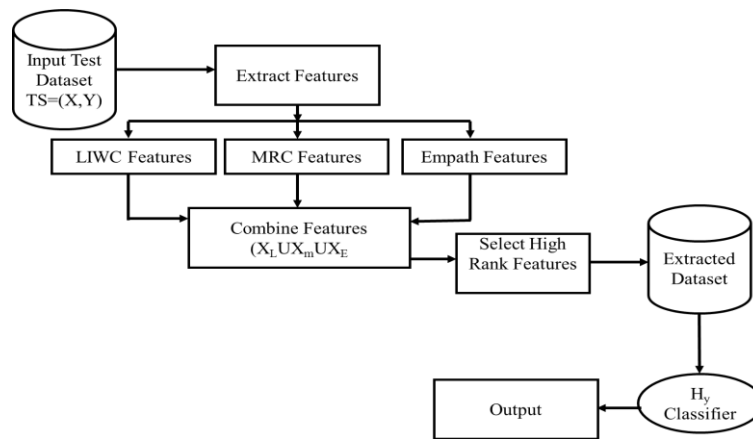


Figure 2. Process of testing model

Algorithm 1. LFSPTI (TR, F, f)

Input parameter-(1) Input Training Dataset $TR = (X, Y)$, where $X \in B^{n \times m}$, $Y \in \{0, 1\}^{n \times 1}$

(2) $P = \{P_1, P_2, P_3, \dots, P_m\}$ is the feature set.

(3) f is the number of features for every feature list and F is the selected relevant number of features.

Output-The best selected feature subset P_F for $y = 1, 2, 3, \dots, l$

1. X_L ← LIWC features from X

```

2.  $X_M \leftarrow$  MRC features from  $X$ 
3.  $X_E \leftarrow$  EMPATH features from  $X$ 
4.  $S \leftarrow$  combine LIWC, MRC and EMPATH such as  $(X_L \cup X_M \cup X_E)$ 
5.  $P_F \leftarrow$  null
6. For  $y \leftarrow 1$  to  $l$  // for every personality trait
  (i)  $P_{F-sta} \leftarrow$  null,  $P_{Chi} \leftarrow$  null,  $P_{MI} \leftarrow$  null and  $P_{PCA} \leftarrow$  null
  (ii) For  $x \leftarrow 1$  to  $m$ 
    (a)  $P_{F-sta}[x] \leftarrow S_x^{(y)}$  // compute F-statistic using equation (5)
    (b)  $P_{Chi}[x] \leftarrow A_x^{(y)}$  // compute Chi-square using equation (6)
    (c)  $P_{MI}[x] \leftarrow MI_x^{(y)}$  // compute Mutual information using equation (7)
    (d)  $P_{PCA}[x] \leftarrow PCA_x^{(y)}$  // compute PCA using PCA method
  (iii) Select high ranked  $f$  features from  $P_{F-sta}$ ,  $P_{Chi}$ ,  $P_{MI}$  and  $P_{PCA}$  using genetic algorithm (see in Algorithm A2) and save in  $P_{F-sta}(HR)$ ,  $P_{Chi}(HR)$ ,  $P_{MI}(HR)$  and  $P_{PCA}(HR)$ , separately.
  (iv)  $P_{temp} \leftarrow$  null,  $count \leftarrow$  null, and  $accuracy \leftarrow 0$ 
  (v) For  $k \leftarrow 1$  to  $F$ 
    (a)  $Pool \leftarrow \{P_{F-sta}(HR)[k], P_{Chi}(HR)[k], P_{MI}(HR)[k] \text{ and } P_{PCA}(HR)[k]\}$ 
    (b) while  $Pool \leftarrow$  not null
      1-  $P \leftarrow$  select and delete a random feature from pool
      2- Train a classifier applying  $P_{temp} \cup P$  as feature set and get its accuracy as  $accuracy_1$ 
      3- If  $accuracy_1 > accuracy$  then:  $P_{temp} = P_{temp} \cup P$  and  $count$  increment ( $count++$ )
      4- If  $count = F$  then: break
  (vi)  $H_y : S_{P_{temp} \rightarrow Y_y}$  // train a classifier for the  $y^{th}$  personality
  (vii)  $P_F[y] = P_{temp}$  // Subset of features for the  $y^{th}$  personality
7. Return  $P_F$ 

```

2.4. Essays dataset and evaluation

Our investigation focuses on Essays Dataset [15], which is the benchmark dataset for personality identification from textual data. It has 2,468 stream-of-consciousness essays. The breakdown of the Essay dataset in terms of personality traits is given in Table 2.

Table 2. The breakdown of the Essay dataset

S. No.	Binary value	Extraversion (EXT)	Neuroticism (NEU)	Agreeableness (AGR)	Conscientiousness (CON)	Openness (OPN)
1.	0	1191	1234	1157	1214	1196
2.	1	1276	1233	1310	1253	1271

As per all the previous studies, we selected accuracy as the prime metric for evaluating the performance of the classifier model. The accuracy for the y^{th} personality trait is computed by (10).

$$Accuracy_y = \frac{\sum_{r \in TS} \|H_y(r) - V_y\|_1}{t} \tag{10}$$

Where TS the test dataset, t is the number of examples in TS, r is a test instance, and V_y is actual value for the y^{th} personality trait. Here, the higher the accuracy value, the better the classification performance.

3. RESULTS AND DISCUSSION

In this part, we express the experiment executed to judge the proposed technique and its comparison made with related existing state-of-the-art techniques as given below:

- In the Mairsee Criterion [20], the SVM classifier trains with all the MRC and LIWC features.
- In FineEmo [22] technique, the SVM classifier works with the Mairsee criterion feature, emotion-based hash tag lexicon feature, and various configurations like clustering-based lexicon features.
- The LIWC and FR [23] technique applies various classifiers with selected features with information gain, PCA, and the entire features.
- In CNN and Mairsee [24], four distinct layouts were used: 1) multilayer perceptron (MLP); 2) MLP plus entirely connected (EC) layer; 3) MLP plus maxpooling layer; 4) MLP plus CNN filter.
- The PTLFM technique [1] uses two classifiers such as SVM and LR, including a balanced weighing mechanism.

Our proposed technique uses three classifiers, SVM, LR, and DT, based on a genetic selection of features. For SVM, the parameters are regularization parameter $C=1.0$ with l_2 squared penalty, radial basis function kernel with three degrees, and tolerance=0.001. For LR, the parameters are regularization parameter

C=1.0 with 12 penalty, tolerance=0.0001, and class weight as balanced. For DT, the parameters are as follows: Criterion as 'gini,' splitter as 'random,' minimum samples at the leaf as one, and class_weight as balanced. Rests of the parameters are default as given in [27]. For the GA, we kept the size of the population at 50, the probability for mutation at 0.01, and the maximum number of generations at 100. The dataset was divided using 5-fold cross validation and experiments were repeated 10 times and the average values are reported.

Table 3 represents the experimental results for accuracy on the Essays dataset. It is evident from the table that the LFSPTI method performs better than all the other existing techniques. The best results for all the other competing methods are 59.26, 61.32, 59.06, 58.94, and 63.03 for EXT, NEU, AGR, CON, and OPN, respectively, whereas the LFSPTI method gives 59.96, 61.83, 60.01, 59.62, and 64.18 for EXT, NEU, AGR, CON, and OPN, respectively. The improvement in classification performance against the best over the competing methods is 1.18%, 0.83%, 1.61%, 1.15%, and 1.82%, for EXT, NEU, AGR, CON, and OPN, respectively. Conceptually, the more information given to the machine learning model, the better its performance; however, if the information given is not helpful, it certainly degrades the performance of the classifiers. This is very true for input feature space. Hence, LFSPTI removes the unuseful features and provides the feature in a way that transforms the input dataset to be very informative for prediction purposes.

Table 3. Personality classification outcomes in terms of accuracy for distinct techniques

Technique	EXT	NEU	AGR	CON	OPN	Mean
Mairesee + SVM	55.49	58.43	55.56	55.27	60.10	56.97
FineEmo + SVM	56.45	58.33	56.03	56.73	60.68	57.64
LIWC and FR + various classifiers	55.75	58.31	57.54	56.04	61.95	57.92
CNN and Mairesee + MLP	55.54	58.42	55.40	56.30	62.68	57.67
CNN and Mairesee + MLP + EC	54.61	57.81	55.84	57.30	62.13	57.54
CNN and Mairesee + MLP + maxpooling	58.09	57.33	56.71	56.71	60.13	57.99
CNN and Mairesee + MLP + CNN filters	55.07	59.38	55.08	55.14	60.51	57.04
PTLFM with SVM	58.85	61.32	58.85	58.85	62.90	60.15
PTLFM with LR	59.26	60.80	59.06	58.94	63.03	60.22
Superior result in all of the above techniques	59.26	61.32	59.06	58.94	63.03	60.22
LFSPTI with SVM	58.97	61.53	58.89	58.91	63.10	60.28
LFSPTI with LR	59.96	61.83	60.01	59.62	64.18	61.12
LFSPTI with DT	58.19	59.88	60.01	59.02	62.88	59.99

This study investigated the effects of MRC, EMPATH, LIWC features, F-statistics, chi-square, MI, and PCA. In contrast, the earlier studies have explored the impact of only a few of these. They have not explicitly addressed the influence of all the combined approaches. Also, the results demonstrate that a good feature selection technique can potentially increase the predictive performance.

We not only compare the performance on individual personality traits but also on the mean value over all the personality traits. The best mean value for accuracy across all the traits for competing methods is 60.28, whereas LFSPTI gives an accuracy of 60.62. The improvement in classification performance against the best over the competing methods is 1.39%. These results demonstrate that LFSPTI gives superior outcomes across all the personality traits compared to all the competing techniques. The outcomes verify the usefulness of LFSPTI as a genetic-based feature selection technique for improving the classification of personality traits. Our study demonstrates that the GA, being evolutionary, is more resilient than the other competing approaches; hence, using better evolutionary algorithms like memetic algorithms in future work may improve it further.

4. CONCLUSION




This paper addresses the core challenge of the low accuracy in predicting personality traits using text processing. The proposed LFSPTI method has solved this problem with the big-five personality trait model from text data, which applies a feature selection method to increase accuracy. LFSPTI gives excellent absolute and comparative enhancement compared to popular methods, in which more than 90% of the features are rejected or removed. The investigation results show that the presented LFSPTI method improves the classification accuracy against the best of the competing methods for all the personality traits and mean overall. The reason for the improvements in the prediction performance is that different parameters provide different aspects of the dataset, and selecting features using genetic selection gives us the most valuable features from the combination of these aspects. Due to the high amount of information and the small number of features, the method has become stable and straightforward. In the future, we plan to use a memetic

algorithm to select the best features instead of genetic selection. Further, some other aspects of the dataset may be considered using other feature selection techniques. Also, datasets other than texts may be considered, where the number of instances for each trait varies heavily for different personality traits to check the performance of the LFSPTI method.




REFERENCES

- [1] N. K. Mishra, A. Singh, and P. K. Singh, "Multi-label personality trait identification from text," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 21503–21519, Mar. 2022, doi: 10.1007/s11042-022-12548-1.
- [2] W. Kang, F. Steffens, S. Pineda, K. Widuch, and A. Malvaso, "Personality traits and dimensions of mental health," *Scientific Reports*, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-33996-1.
- [3] R. Divi and C. S. Potala, "Correlation based data unification for personality trait prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 404–411, Jan. 2023, doi: 10.11591/ijeecs.v29.i1.pp404-411.
- [4] M. Salahat, L. Ali, T. M. Ghazal, and H. M. Alzoubi, "Personality assessment based on natural stream of thoughts empowered with machine learning," *Computers, Materials and Continua*, vol. 76, no. 1, pp. 1–17, 2023, doi: 10.32604/cmc.2023.036019.
- [5] I. B. Myers, "Mbt manual: a guide to the development and use of the myers-briggs type indicator," *Consulting Psychologists Press, Palo Alto*, 1998.
- [6] R. Singh *et al.*, "Antisocial Behavior identification from twitter feeds using traditional machine learning algorithms and deep learning," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 4, pp. 1–17, May 2023, doi: 10.4108/eetsis.v10i3.3184.
- [7] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2313–2339, Oct. 2020, doi: 10.1007/s10462-019-09770-z.
- [8] G. Jiang, M. Xu, S. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and inducing personality in pre-trained language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [9] R. Arya, J. Singh, and A. Kumar, "A survey of multidisciplinary domains contributing to affective computing," *Computer Science Review*, vol. 40, p. 100399, May 2021, doi: 10.1016/j.cosrev.2021.100399.
- [10] J. Serrano-Guerrero, B. Alshouha, M. Bani-Doumi, F. Chiclana, F. P. Romero, and J. A. Olivias, "Combining machine learning algorithms for personality trait prediction," *Egyptian Informatics Journal*, vol. 25, p. 100439, Mar. 2024, doi: 10.1016/j.eij.2024.100439.
- [11] N. K. Mishra and P. K. Singh, "FS-MLC: Feature selection for multi-label classification using clustering in feature space," *Information Processing and Management*, vol. 57, no. 4, p. 102240, Jul. 2020, doi: 10.1016/j.ipm.2020.102240.
- [12] M. D. Kamalesh and B. B, "Personality prediction model for social media using machine learning Technique," *Computers and Electrical Engineering*, vol. 100, p. 107852, May 2022, doi: 10.1016/j.compeleceng.2022.107852.
- [13] G. Farnadi, S. Zoghbi, M. F. Moens, and M. De Cock, "How well do your Facebook status updates express your personality?," *In 22nd Edition of the Annual Belgian-Dutch Conference on Machine Learning (BENELEARN)*, p. 88, 2013.
- [14] A. Bruno and G. Singh, "Personality traits prediction from text via machine learning," in *Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022*, Jun. 2022, pp. 588–594, doi: 10.1109/AIC55036.2022.9848937.
- [15] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference," *Journal of Personality and Social Psychology*, vol. 77, no. 6, pp. 1296–1312, 1999, doi: 10.1037/0022-3514.77.6.1296.
- [16] R. K. Cherukuru, A. Kumar, S. Srivastava, and V. K. Verma, "Prediction of personality trait using machine learning on online texts," in *2022 International Conference for Advancement in Technology, ICONAT 2022*, Jan. 2022, pp. 1–8, doi: 10.1109/ICONAT53423.2022.9725910.
- [17] H. A. Schwartz *et al.*, "Personality, gender, and age in the language of social media: the open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, p. e73791, Sep. 2013, doi: 10.1371/journal.pone.0073791.
- [18] R. L. Vásquez and J. Ochoa-Luna, "Transformer-based approaches for personality detection using the MBTI model," in *Proceedings - 2021 47th Latin American Computing Conference, CLEI 2021*, Oct. 2021, pp. 1–7, doi: 10.1109/CLEI53233.2021.9640012.
- [19] A. Wijaya, I. Prasetya, N. Febrianto, and D. Suhartono, "The 'big five traits' personality prediction system from twitter data, [Translation] (in Indonesia: Sistem prediksi kepribadian 'the big five traits' dari data Twitter)," *Bina Nusantara University*, 2016.
- [20] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, Nov. 2007, doi: 10.1613/jair.2349.
- [21] M. Coltheart, "The mrc psycholinguistic database," *The Quarterly Journal of Experimental Psychology Section A*, vol. 33, no. 4, pp. 497–505, Nov. 1981, doi: 10.1080/14640748108400805.
- [22] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, Jan. 2015, doi: 10.1111/coin.12024.
- [23] E. Tighe, J. Ureta, B. Pollo, C. Cheng, and R. D. D. Bulos, "Personality trait classification of essays with the application of feature reduction," in *Proceedings of the 4th workshop on Sentiment Analysis where AI meets Psychology (SAAIP)*, pp. 22–28, 2016.
- [24] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, Mar. 2017, doi: 10.1109/MIS.2017.23.
- [25] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Lawrence Erlbaum Associates, Mahway*, 2001.
- [26] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Conference on Human Factors in Computing Systems - Proceedings*, May 2016, pp. 4647–4657, doi: 10.1145/2858036.2858535.
- [27] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, pp. 2825–2830, 2011.




BIOGRAPHIES OF AUTHORS

Angad Singh    is pursuing Ph.D. in Computer Science and Engineering from Rabindranath Tagore University, Bhopal, India. He completed his M.Tech from Shri Ram Institute of technology, Jabalpur, India in 2011. He has more than 16 years of teaching experience in reputed institutions like Bansal Institute of Science and Technology, Bhopal. His area of interested is machine learning and deep learning. He can be contacted at email: angada2007@gmail.com.






Priti Maheshwary    is Professor in Computer Science and Engineering Department at Rabindranath Tagore University, Bhopal. She had her Doctorate from MANIT, Bhopal in Remote Sensing Image Retrieval. She has completed research project on Climate Change detection and monitoring funded by SAC and Environment Monitoring using Sensor devices funded by AISECT University. Ongoing Project is on Crop Monitoring sponsored by SAC. She is the author of more than 20 publications in different journals of repute. She is the author of book chapter on Software Copyright. Her interests include machine learning, cyber physical systems, mobile networks, WSN, adhoc networks, data mining, and image processing. Her work experience includes 20 years in teaching computer science and engineering. She is guiding Ph.D. in the field of machine learning, IoT and networks. She can be contacted at email: pritimaheshwary@gmail.com.



Nitin Kumar Mishra    is a Senior Assistant Professor at the School of Computing Science and Engineering, VIT-Bhopal University, India. He has a Doctorate in Information Technology from Atal Bihari Vajpayee-Indian Institute of Information Technology and Management (ABV-IIITM), Gwalior. He has 16 years of academic experience. As a researcher, designing, developing, and optimizing machine learning algorithms for multi-label classification problems is his area. He has presented and published many research papers in SCIE/Scopus-indexed international conferences and journals. He reviews various journals such as Information Sciences, Information Processing and Management, Computers and Electrical Engineering, Expert System with Applications, and Artificial Intelligence Review. He can be contacted at email: nkmishra0701@gmail.com.



Timothy Malche    is an Assistant Professor in Computer Science and Applications Department at Manipal University, Jaipur, India. He has a Doctorate in Computer Science and Applications from Rabindranath Tagore University, Bhopal. He has 18 years of academic experience and Computer Specialist with a demonstrated history of working in the higher education industry. Skilled in research, lecturing, technical writing, teaching, and computer programming. Strong information technology professional with a MCA+Ph.D. He can be contacted at email: timothy.malche@jaipur.manipal.edu.