Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12548-1

Multi-label personality trait identification from text

Nitin Kumar Mishra[1] · **Aditya Singh**[2] · **Pramod Kumar Singh**[1]

Received: 26 July 2021 / Revised: 29 November 2021 / Accepted: 31 January 2022

Abstract

Understanding the personality is beneficial for many purposes, e.g., it is natural to predict a user's personality before offering him or her any services. The personality is intrinsic in the behavior of a person in all aspects, such as text writing. Some work has been proposed in recent times for correctly classifying a person's personality from the text. However, it is still a significant challenge as the achieved accuracy is low; therefore, the proposed work addresses this issue. Effective feature selection techniques provide better classification accuracy in multi-label classification and personality traits identification as multi-label classification problem requires efficacy of feature selection methods. Therefore, to improve the accuracy using feature selection technique, this paper proposes a method for personality trait recognition from textual data called *P*ersonality *T*rait Classification based on *L*inguistic and *F*eature selection as *M*ulti-label classification (*PTLFM*). It combines analysis of variance's F-statistic, Chi-square, and Mutual information with the sequential feature selection wrapper method to rank features. These three criteria apprehend different aspects of the dataset. The experimental results demonstrate that the proposed *PTLFM* method achieves higher accuracy across all the personality traits than the prevailing state-of-the-art machine learning and deep learning models. PTLFM provides an impressive absolute improvement of 2.23% and 3.84% of comparative improvement over the existing prevalent method, with more than 90% of features discarded. Furthemore, the proposed PTLFM achieves a percentage gain compared to the competitive methods across different personality traits Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness in absolute terms 1.17, 1.94, 2.35, 1.64, and 0.35 respectively, and in comparative terms 2.01, 3.27, 4.14, 2.86, and 0.56 respectively. The results suggest that although deep learning is a popular paradigm, it does not always lead to a better predictive performance than machine learning models in all the problem domains.

Keywords Personality trait identification · Feature selection · ANOVA's F-statistic · Chi-square · Mutual information · Multi-label classification

✉ Nitin Kumar Mishra
nkmishra0701@gmail.com

Extended author information available on the last page of the article.

1 Introduction

Personality represents multiple attitudes of a person, such as behavior, feeling and thought [14]. A person's personality is a permanent set of traits and styles that he or she exhibits. It is a set of characteristics representing personal inclination and distinguishes him or her from the "normative normal person" in his or her society. Here, the trait is an engagement in a specific type of behavior and experiencing individual emotional states. Styles is the disposition of having to do, i.e., how to do and not about what to do [4]. Personality is labeled with two widely used models: The big-Five model and Myers-Briggs Type Indicator (MBTI) model. The Big-five model [11] uses personality with five traits: Insurgency vs. Extraversion (EXT), Disagreeable vs. Agreeableness (AGR), Unconsciously vs. Conscientiousness (CON), Emotional Stability vs. Neuroticism (NEU), and Openness to Experience (OPN). The MBTI model [23] uses four personality traits, namely extroversion (E) vs. introversion (I), sensing (S) vs. intuition (N), thinking (T) vs. feeling (F), and Judging (J) vs. perceiving (P). However, the Big-Five model is a standard model due to its formal language description of personality for machine learning and deep learning models [3, 17].

In both of the above models, the personality traits are binary variables. Based on this binarization, the traits may be classified as 32 classes in the Big-five model because the binarization of five digits gives the number 32 and 16 classes in the MBTI model because the binarization of four digits gives the number 16. This way, it becomes a multi-class problem [39]. However, the multi-class assumption introduces a higher imbalance due to the higher number of classes. Figure 1 represents the frequency distribution of the Essays dataset [25], which is prevalent and used in this paper's experimentation also. The minimum number of relevant instances for class 2 is 43, and the maximum number of relevant instances for class 23 is 189. The difference between the relevant number of exemplars for different classes creates a significant imbalance in the Essays dataset. Further, it has to train quite a significant number of classifiers for the multi-class problem in one versus one and even in one versus rest [28, 34]. Therefore researchers have considered it as a multi-label problem, and we are also doing so.

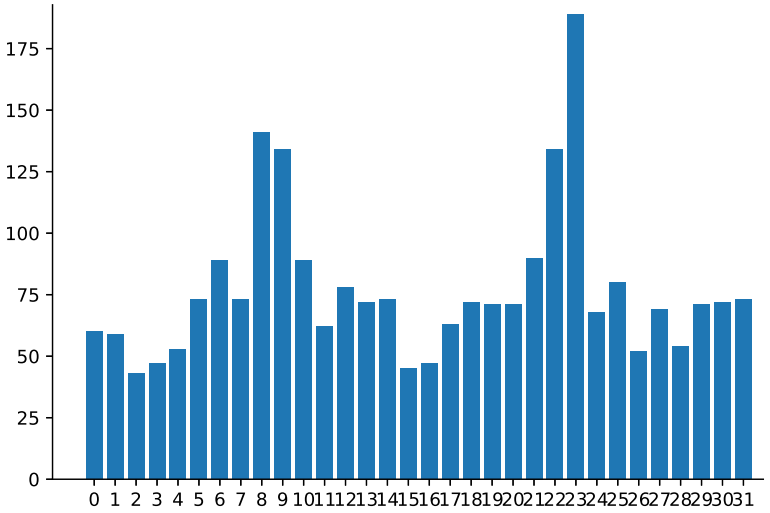

Fig. 1 Frequency distribution classes of Essays dataset as multi-class dataset

Multi-label classification involves the exemplars associated with more than one label simultaneously. As discussed above, the Big-five model of personality has five different traits. It may tag an individual with more than one label simultaneously, e.g., an individual may be categorized as Agreeable, Intellect, and Extravert. Multi-label classification algorithms are categorized as first-order, second-order, or high-order depending on the considerations of no label-interdependence, pairwise-interdependence, or dependence among three or more labels, respectively [20]. Since the personality traits have a very low correlation among them, nearly all personality trait identification methods treat the problem as having no dependence among labels.

Personality trait identification has many applications, e.g., personality traits affect the decision-making process of an individual, and it helps managers to understand an employee's behavior for specific situations [12]; based on personality traits different employees may be used in different project phase to complete it efficiently [6, 35]; marketing strategy for consumer goods or tourist places may be based on personality traits [1, 5, 32].

Personality trait identification is envisioned as a promising approach to address challenges in all the above applications. However, the trait identification approaches witness a low accuracy score. Various machine learning and deep learning models are proposed recently to overcome this limitation. Nearly all the models use textual features such as 1-gram, 2-gram, and n-gram or linguistic features such as word count, number of pronouns, and number of syllables, or even a combination of both text features and linguistic features. Rather than relying on a full feature set, the prediction performance can be increased by feature selection methods [19]. Therefore, this paper proposes a feature selection method based on ANOVA's F-statistic, Chi-square, and Mutual information to address this challenge, achieving higher accuracy than the prevalent personality trait identification methods. The main contributions of this work are as follows.

- This paper proposes a novel method called *P*ersonality *T*rait Classification based on *L*inguistic and *F*eature selection as *M*ulti-label classification (*PTLFM*) for feature extraction.
- *PTLFM* uses high ranked features out of 102 LIWC and MRC features obtained from the input dataset. For ranking and selecting features, it combines ANOVA's F-statistic, Chi-square and Mutual information with sequential feature selection wrapper approach.
- PTLFM first uses three filter methods to reduce the number of features and applies a wrapper-based method to the reduced feature set further; it is not done previously in the literature for personality trait identification.
- It trains machine learning models Support Vector Machine (SVM) and Logistic Regression (LR) on the selected features.
- The experimental results, on the well-known *Essays* dataset, show that the method performs superior compared to other state-of-the-art prevailing machine learning and deep learning methods.
- The results puts an emphasis on the capability of machine learning method that may outperform a deep learning method. Deep learning is a prevalent paradigm; however, it may not surpass the machine learning methods in all the problem domains, specifically when the dataset is simple and we can obtain a feature set of manageable size.

To the best of our knowledge, this is the first work that combines LIWC and MRC features and utilizes F-statistic, Chi-square, and Mutual information with a wrapper-based sequential feature selection approach to select features for personality trait identification.

The rest of the paper is organized as follows. Section 2 presents work related to personality trait identification from textual data. Sections 3 and 4 describe the proposed method with related concepts and the dataset with performance metrics used in the experimentation. Section 5 presents results and discussion. Finally, Section 6 provides the conclusion and future research directions.

2 Related work

This section presents the work related to personality trait identification that consider it as a multi-label problem. The Linguistic Styles [25] was the first pioneered work that shows written language is unique for each person, and personality traits can be identified using language cues. It is a meaningful and independent means of examining personality. Linguistic Inquiry Word Count (LIWC) tool [24] is a text analysis tool that calculates the frequency for different categories of words in a given text. The authors find notable correlations between the linguistic dimensions and personality traits, e.g., Neurotics express more negative emotion words and the first person singular pronouns, and fewer positive emotion words. On the contrary, an agreeable person is more inclined to use optimism and less willing to use negative emotions and articles.

The Mairesse baseline [15] performs experiments with LIWC for Big-five personality traits using both conversational and textual dataset. For the textual dataset, Mairesse baseline relies on correlations between linguistic features and personality traits. It extracted a total of 102 features with 88 Linguistic Inquiry and Word Count (LIWC) features from the LIWC utility and 14 MRC Psycholinguistic database (MRCPD) [7] features from MRCPD. The initial version of MRCPD contains statistics for 98,538 words, and the current version contains statistics for 150,837 words. It represents some of the relevant features for each word out of 26 different linguistic features such as letters-count, phonemes-count, syllables-count, Familiarity, Concreteness, Varient Phoneme, Written Capitalised, Irregular Plural, and Stress Pattern.

FineEmo [22] is a hashtag emotion corpus for six emotions: joy, anger, fear, sadness, surprise, and disgust. It is created from a corpus of 21,000 tweets extracted by authors. The FineEmo uses word-based emotion associations to improve the personality prediction accuracy on the Essays dataset, assuming that emotions have a strong relationship with personality traits. The hashtag lexicon features are also combined with Mairesse features that result in some improvement in results.

LIWC+FR [33] is a personality trait identification method utilizing LIWC features and feature reduction. It uses Principle Component Analysis (PCA) and Information Gain (IG) for feature reduction and Support Vector Machine and Logistic regression as a classifier. It shows that dimensionality reduction decreases the time complexity and increases the accuracy of the classifiers. However, it utilizes the LIWC feature set only; adding some other features may increase the classification accuracy.

The CNN+multichannel [13] was the first model for text classification that utilized a convolution neural network (CNN) with word2vec vectorization [18]. First, it converts each word into a word vector and then applies sentence vectorization on top of word2vec conversion of the words in a sentence. The CNN+Mairesse [16] is based on CNN+multichannel that uses Big-five model of personality traits on the Essays dataset. First, Word2vec embedding is used to code each word in a 300-dimensional vector. Then it uses sentence-level and document-level vectorization. Further, a CNN based classifier is used to extract 600 features. These features are combined with 84 Mairesse features, creating a 684 featured vector

for each document. Finally, a fully connected multi-layer perceptron with two softmax neurons is used. It trains different classifiers for each personality trait. The AttLSTM [39] is a model that uses attention mechanism with LSTM (Long Short Term Memory).

The Vertex-weighted Multi-modal Multitask Hypergraph Learning (VM2-HL) [40] model learns perceived emotion recognition from physiological signals. Since personality influences the perceived behavior, VM2HL incorporates personality into the multi-modal physiological features. It uses a hypergraph, which uses the relationship between physiological signals and personality and gives different weights to hyperedges. The fusion of personality with physiological feature improves the recognition of emotions. Also, user interest mining can benefit from incorporating the user's personality, which might infer the topical interests [8]. The authors have incorporated Big-five personality traits in the dataset using linguistic features and shown that including users' personalities in interest mining improves the system's prediction accuracy.

A comparison of features used in different competing methods is shown in Table 1. Conceptually, all the above methods are analogous: They all use the full set of features, i.e., LIWC only or LIWC and MRC combined. The LIWC+FR uses dimensionality reduction; however, it utilizes only the LIWC features. To address this issue, we propose PTLFM, in contrast, which utilizes all the LIWC and MRC features and applies feature selection using F-statistic, Chi-square and Mutual information with sequential feature selection wrapper method. The PTLFM aims to improve accuracy using different sets of features for each personality trait.

A single feature selection filter method may not select the best possible feature subset as it focuses on a particular aspect of the training data. Hence, we use three different feature selection filter methods focusing on different aspects of the dataset. As these three methods provide different and divergent feature subsets, simply making a union of these three feature subsets creates a more extensive feature set. To keep the feature subset of manageable size, we further use a wrapper method that selects the high-ranked features out of the feature subsets determined by the filter method; this way, PTLFM is a hybrid of filter and wrapper method.

There are various methods for trait identification beyond text usage that utilizes feature reduction. Wang et al. [36] identifies the Big-five traits from the drivers' signals and uses PCA and sequential backward feature selection method for feature reduction. Pohjalainen et al. [26] performs a speaker's Big-five personality trait classification form recorded speech and suggests that a combination of different feature subsets may improve the performance of classifiers. Al Marouf et al. [2] applies Information gain, chi-square, and correlation to select high ranked features. Mishra et al. [21] performs personality traits identification form

Table 1 Features for different methods

Title	No. of features	Classification model	Performance metric (Results)
Mairesse baseline [15]	102 (LIWC+MRC)	SVM	Accuracy (56.97)
FineEmo [22]	102(LIWC+MRC)+ 585 (hashtag lexicon)	SVM	Accuracy (57.64)
LIWC+FR [33]	56 (PCA), 10 (IG)	SVM, LR	Accuracy (57.92)
CNN+Mairesse [16]	600(CNN)+84(LIWC+MRC)	MLP	Accuracy (57.99)

an audio-visual recording of interviewers and applies feature selection using mutual information prosodic non-verbal features. Identification of Big-five personality traits is presented in [10] and [29] using deep learning methods as image processing. Tayarani et al. [31] identifies Big-five personality traits from audio clips of speeches by utilizing the fillers, i.e., small utterance of the speakers.

Xue et al. [37] utilizes a two-layered Recurrent Convolution Neural Network with an attention mechanism for personality trait identification of social media users. The dataset used is myPersonality. El-Demerdash et al. [9] fuses the different datasets at data level and employs various pre-trained deep learning model for personality trait classification.

3 Personality trait classification based on linguistic and feature selection as multi-label classification (*PTLFM*)

This section describes the proposed *PTLFM* method, which is dependent on LIWC and MRC features, and ANOVA's F-statistic, Chi-square, and Mutual information. Hence, the first two subsections elaborate on these concepts, and the final subsection presents the proposed method.

Formally, as a multi-label classification, the personality trait identification problem can be represented as follows. Given a training dataset $D = (X, Y)$, $X \in R^{n \times m}$ and $Y \in \{0, 1\}^{n \times l}$, where n is the number of exemplar, m is the number of features, and l is the number of different personality class labels. The objective of the personality trait identification problem is to learn a classifier h form D such that for an unknown instance x_u, $h(x_u)$ can predict its personality traits as close as the ground truth traits. For Big-five personality trait, $l = 5$.

Problem transformation (PT) and Algorithm adaptation (AA) are the two broad categories of MLC algorithms. The PT methods transform the problem so that existing single-label classification algorithms may be applied directly. The AA methods transform the existing single-label classification algorithms in such a way that they can solve the MLC problem. The PT methods are the predominant method in MLC literature. One of the straightforward PT methods is Binary Relevance (BR) that considers the MLC problem as a multiple SLC problem and solves the SLC problem for each label, and finally combines the results of all the SLC problems. BR method is the method of choice when there is little or no correlation among the labels. Figure 2 represents the correlation among the different class labels. It is evident from Fig. 2 that the personality traits of the Essays dataset have little correlation among them. Therefore, this paper uses the identification of different personality traits in a BR fashion.

3.1 LIWC and MRC features

Like the Mairesse baseline, we have extracted LIWC features and MRC features from the LIWC utility and MRC, respectively. We have extracted LIWC features from five different categories, which are as follows. Standard counts: different features such as Word count, words per sentence, and the number of articles; Psychological processes: features such as positive emotions, anger, and certainty; Relativity: features such as the past tense verb, up, and exclusive; Personal concerns: features such as work and job, sports, home, and sleeping; and Other dimensions: features such as comma, colon, semi-colon, and apostrophe. We have also extracted different MRC features such as the number of letters, syllables, phonemes, concreteness, and familiarity rating.

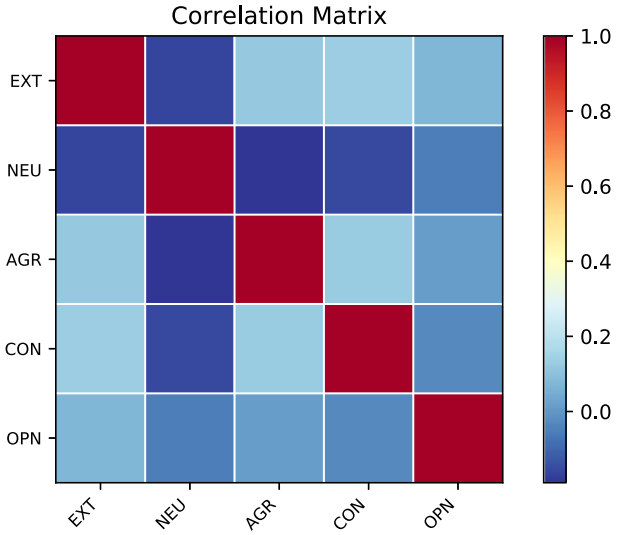

Fig. 2 Correlation among class labels of Essays dataset

3.2 F-statistic, chi-square, and mutual information

To calculate F-statistic for the i^{th} feature relative to the j^{th} personality trait, it selects relevant instances' i^{th} feature using (1) and irrelevant instances' i^{th} feature (2).

$$X_i^{1(j)} = \{x_{ki} : y_{ki} = 1 \; for \; k = 1, 2, ..., n\} \tag{1}$$

$$X_i^{0(j)} = \{x_{ki} : y_{ki} = 0 \; for \; k = 1, 2, ..., n\} \tag{2}$$

Next, the mean of the sets $X_i^{1(j)}$ and $X_i^{0(j)}$ are calculated using (3) and (4), respectively.

$$\mu_1^{(j)} = mean \; over \; X_i^{1(j)} = \frac{\sum_{x \in X_i^{1(j)}} x}{n_1} \tag{3}$$

and

$$\mu_0^{(j)} = mean \; over \; X_i^{0(j)} = \frac{\sum_{x \in X_i^{0(j)}} x}{n_0} \tag{4}$$

where $n_1 = |X_i^{1(j)}|$ and $n_0 = |X_i^{0(j)}|$

Finally, (5) calculates i^{th} feature's F-statistic.

$$F_i^{(j)} = \frac{\left(\mu_0^{(j)} - \mu^{(i)}\right)^2 + \left(\mu_1^{(j)} - \mu^{(i)}\right)^2}{\frac{1}{n_0}\sum_{x \in X_i^{0(j)}} \left(x - \mu_0^{(j)}\right)^2 + \frac{1}{n_1}\sum_{x \in X_i^{1(j)}} \left(x - \mu_1^{(j)}\right)^2} \tag{5}$$

where $\mu^{(i)}$ is the overall mean of X_i.

To calculate Chi-square for the i^{th} feature relative to the j^{th} personality trait, (6) is used.

$$\chi_i^{(j)} = \sum_{c\in\{0,1\}} \frac{\left(X_{obs(i,c)}^{(j)} - X_{exp(i,c)}^{(j)}\right)^2}{X_{exp(i,c)}^{(j)}} \tag{6}$$

where $X_{obs(i,c)}^{(j)}$ and $X_{exp(i,c)}^{(j)}$ are the observed and expected values of i^{th} feature corresponding to the j^{th} personality trait, respectively; c indicates the relevancy of the instances.

Entropy quantifies the disorder present in the data. For a discrete random variable $\mathscr{X}$, it is calculated by (7).

$$Entropy(\mathscr{X}) = -\sum_{i=1}^{n} p(x_i) * log(p(x_i)) \tag{7}$$

where $p(x_i)$ is the probability of x_i over the domain of $\mathscr{X}$ denoted by $\{x_i\}_{i=1}^{i=n}$.

If the two variables are dependent, the entropy of one variable decreases after observing the other. Conditional entropy quantifies the remaining uncertainty of one variable when the other is known. For two random variables $\mathscr{X}$ and $\mathscr{Y}$, it is calculated by (8).

$$cEntropy(\mathscr{X}/\mathscr{Y}) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p(y_j)p(x_i/y_j) * log(p(x_i/y_j)) \tag{8}$$

where $p(x_i/y_j)$ is the conditional probability of x_i given y_j over the domain of $\mathscr{X}$ denoted by $\{x_i\}_{i=1}^{i=n}$ and the domain of $\mathscr{Y}$ denoted by $\{y_j\}_{j=1}^{i=m}$.

Mutual information specifies the gain obtained for one variable concerning the other. It quantifies the dependence of two variables and is calculated using (9).

$$MI(\mathscr{X},\mathscr{Y}) = Entropy(\mathscr{X}) - cEntropy(\mathscr{X}/\mathscr{Y}) \tag{9}$$

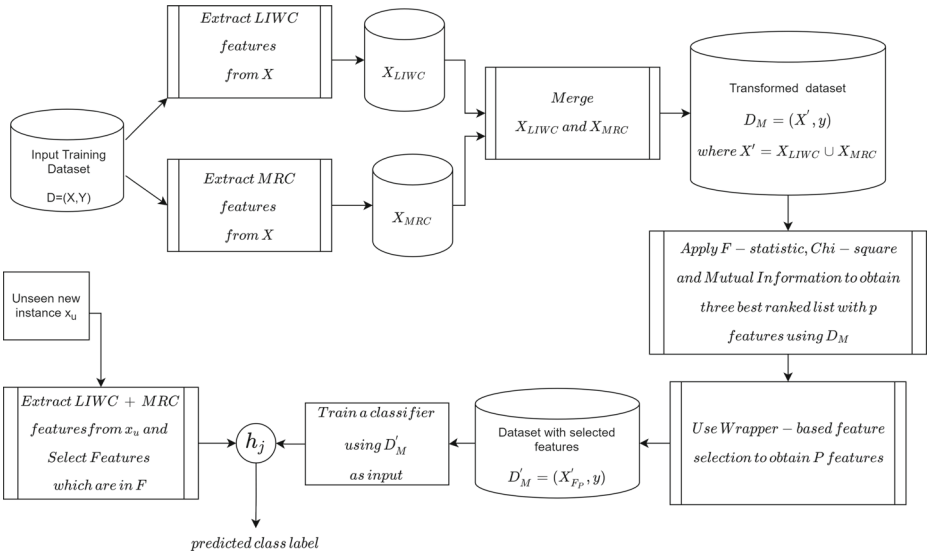

Fig. 3 Block diagram for personality trait identification PTLFM

3.3 Working of the method

The method works in a label-wise manner (refer to Fig. 3 and Algorithm 1) for j^{th} personality trait as follows. First, it finds all the LIWC features X_{LIWC} and MRC features X_{MRC} from the given input dataset D, and merge these features to obtain a single feature set $X' = X_{LIWC} \cup X_{MRC}$. From this combined feature set X', it calculates F-statistic, Chi-square, and Mutual information for each feature.

Algorithm 1 PTLFM_train (D, P, p).

Input: $D = (X, Y)$ be the input dataset, where $X \epsilon R^{n \times m}$, $Y \epsilon \{0, 1\}^{n \times l}$, $F = \{F_1, F_2, \cdots, F_m\}$ is the feature set and $L = \{L_1, L_2, \cdots L_l\}$ is the set of class labels. p is the number of features for each feature list and P is the final number of features selected.

Output: The selected feature subset F_P for j = 1, 2, ..., l

1. $X_{LIWC} \leftarrow$ Extract LIWC features from X
2. $X_{MRC} \leftarrow$ Extract MRC features from X
3. $X' \leftarrow$ Merge X_{LIWC} and X_{MRC}
4. $F_P \leftarrow \phi$
5. For $j \leftarrow 1 \ to \ l$ // for each personality trait
   - a $F_{F-stat} \leftarrow \phi$, $F_{Chi2} \leftarrow \phi$, and $F_{MI} \leftarrow \phi$
   - b For $i \leftarrow 1 \ to \ m$
     - i $F_{F-stat}[i] \leftarrow F_i^{(j)}$ // calculated using (5)
     - ii $F_{Chi2}[i] \leftarrow \chi_i^{(j)}$ // calculated using x(6)
     - iii $F_{MI}[i] \leftarrow MI_i^{(j)}$ // calculated using (9)
   - c Select top-ranked p features from F_{F-stat}, F_{Chi2}, and F_{MI} and store in F^p_{F-stat}, F^p_{Chi2}, and F^p_{MI} , respectively.
   - d $count \leftarrow 0, acc \leftarrow 0, F_{temp} \leftarrow \phi$
   - e For $k \leftarrow 1 \ to \ P$
     - i $pool \leftarrow \{F^p_{F-stat}[k]$, $F^p_{Chi2}[k]$, and $F^p_{MI}[k]\}$
     - ii while *pool* is not empty
       - A $f \leftarrow$ select and remove a random feature from pool
       - B train a classifier using $F_{temp} \cup f$ as feature set and find its accuracy as $acc1$
       - C if $acc1 > acc$ then: $F_{temp} = F_{temp} \cup f$ and $count++$
       - D if *count=P* then: *break*
   - f $h_j : X'_{F_{temp}} \rightarrow Y_j$ // Train a classifier for the j^{th} label
   - g $F_P[j] = F_{temp}$ // subset of features for the j^{th} label
6. return F_P

Algorithm 2 $PTLFM\_predict\,(T, F_P)$.

Input: $T = (X, Y)$ be the input dataset, where $X \epsilon R^{n \times m}$, $Y \epsilon \{0, 1\}^{n \times l}$, $F = \{F_1, F_2, \cdots, F_m\}$ is the feature set and $L = \{L_1, L_2, \cdots L_l\}$ is the set of class labels. F' is the feature subset selected by the Algorithm 1.

Output: The predicted labels for j = 1, 2, ..., l

1. $X_{LIWC} \leftarrow$ Extract LIWC features from X
2. $X_{MRC} \leftarrow$ Extract MRC features from X
3. $X' \leftarrow$ Merge X_{LIWC} and X_{MRC}
4. $Y_{pred} = \phi$
5. For $j \leftarrow 1\ to\ l$
   (a) // Predict the j^{th} label with selected features
   (b) $Y_{pred}[j] \leftarrow h_j(X'_{F_P})$
6. return Y_{pred}

Next, it selects three subsets of p number of features (determined empirically) having highest score of F-statistic, Chi-square and Mutual Information denoted as F^p_{F-stat}, F^p_{Chi2}, and F^p_{MI}, respectively. Further, it randomly selects one feature among the first ranked features from these three lists and trains a classifier. The rationale behind selecting a feature randomly from the feature pool is that we cannot determine which feature selection method delivers the best feature ranking. If a feature is equally good from the perspective of more than one criterion, its chances of getting selected at an earlier stage are higher. The set of selected features is denoted by F_P. Again from the remaining two first-ranked features, it selects one, adds it to F_P, and trains a classifier. If this currently trained classifier has an accuracy lower than the previous classifier, the feature is removed from the feature set F_P. Now, it uses the first-ranked feature from the third list, adds it to F_P, and trains the classifier. A feature sustains in F_P only if its addition to F_P improves the accuracy of the classifier. The same procedure is repeated for the second, third, and other ranked features from all the lists until all the lists are exhausted, or P number of features are eventually selected. Kindly note that a feature is added to F_P only if it is already not in F_P. Finally, *PTLFM* trains a classifier using the selected feature set, i.e. X'_{F_P}.

For predicting (refer to Algorithm 2) personality trait of an unseen instance x_u, it extracts LIWC, and MRC features, combine them into a single feature set and passes to the classifier only the selected features specified by F_P. To find the optimal number of features, we have used the grid search method using accuracy measure with internal cross-validation. The number of features is selected from the set {6, 7, 8, 9}, as suggested in [30] and [38]. We kept $p = P$ for the sake of simplicity and it also makes the analysis of the experimentation simple in Section 5.1.

4 Dataset and performance metric

The Essays dataset [25] is used in the experimentation that contains 2,467 anonymous stream-of-consciousness essays, where authors' personality traits are tagged using Big-Five personality traits. The distribution of relevant exemplars for EXT, NEU, AGR, CON,

Table 2 Distribution of Essays dataset

	EXT	NEU	AGR	CON	OPN	Total
Number of relevant examplars	1276	1233	1310	1253	1271	6343
Number of irrrelevant examplars	1191	1234	1157	1214	1196	5992
Fraction of relevant over total exemplars	0.517	0.5	0.531	0.508	0.515	0.514

and OPN personality traits are shown in Table 2. It is evident from the fraction of relevant exemplar over total exemplar that the dataset is quite balanced across all personality traits. Further, the number of relevant labels per exemplar, also called density, is 0.514, and it indicates that the dataset is also neither dense nor sparse. Another Big-five personality trait social science text dataset is MyPersonality dataset [27], which is created by a Facebook App called myPersonality, in which Facebook users participated by filling a psycho-linguistics questionnaire. However, it is currently unavailable as its creator has stopped providing it outside their research lab after May 2018.

Since the dataset is balanced in all aspects, we have used personality trait-wise accuracy to evaluate the proposed method's performance and compare it against other methods. For an individual personality trait, classification accuracy is the fraction of exemplars correctly classified by a classifier over all the given exemplars. For the j_{th} personality trait it is calculated by (10).

$$Acc_j = \frac{\sum_{x_u \in T} ||h_j(x_u) = y_{uj}||_1}{q} \tag{10}$$

Here, T is the test dataset containing q number of test exemplars, y_{uj} is the ground truth for j^{th} personality trait of the test exemplar x_u. For any personality trait, higher the value of accuracy better is the classification performance.

To identify the personality of a person all the personality traits need to be identified simultaneously. With this perspective a strict performance metric is Exact match, which considers a classification correct only if all the predicted traits of a person matches with the ground truth, otherwise it is an incorrect classification. The Exact-match is calculated using the (11).

$$Exact - match = \frac{\sum_{x_u \in T} \left(\prod_{j=1}^{l} ||h_j(x_u) = y_{uj}||_1 \right)}{q} \tag{11}$$

5 Experimentation and discussion

This section describes the experimentation performed to evaluate the proposed method and compare it with other state-of-the-art related methods, namely Mairesse baseline, FineEmo, LIWC+FR, and CNN+MAiresse. The Mairsse baseline uses all the LIWC and MRC features and trains an SVM classifier. The FineEmo uses an SVM classifier with different configurations such as cluster lexicon features, Mairesse baseline feature, and emotion hashtag lexicon. The LIWC+FR uses different classifiers using full feature set, PCA feature set and IG feature set. The CNN+Mairesse uses four different configurations: Multilayer perceptron (MLP), MLP with a fully connected layer, MLP with MaxPooling and MLP with convolution filters. The proposed PTLFM uses SVM and Logistic Regression (LR) with a balanced weighing scheme. All these methods and our proposed PTLFM method use 10-fold cross-validation, and average results are reported across all the personality traits. The

Table 3 Personality trait wise Accuracy for different methods

Method	Personality traits					
	EXT	NEU	AGR	CON	OPN	Average
Mairesse Baseline with SVM	55.49	58.43	55.56	55.27	60.10	56.97
FineEmo	56.45	58.33	56.03	56.73	60.68	57.64
LIWC + FR	55.75	58.31	57.54	56.04	61.95	57.92
AttLSTM	55.79	59.23	55.18	57.14	61.95	57.86
CNN + Mairesse (MLP+FC)	54.61	57.81	55.84	57.30	62.13	57.54
CNN + Mairesse (MLP+MP)	58.09	57.33	56.71	56.71	61.13	57.99
CNN + Mairesse (MLP)	55.54	58.42	55.40	56.30	62.68	57.67
CNN + Mairesse (MLP+2,3,4 filters)	55.07	59.38	55.08	55.14	60.51	57.04
Best for All of the above	58.09	59.38	56.71	57.30	62.68	57.99
PTLFM + LR	**59.26**	60.80	**59.06**	**58.94**	**63.03**	**60.22**
PTLFM + SVM	58.85	**61.32**	58.85	58.85	62.90	60.15

FC: Fully Connected layer, MP: Maxpooling, MLP: Multilayer Perceptron

dataset is normalized using z-score normalization. Table 3 reports the best results over all the configurations for FineEmo and LIWC+FR.

From the Table 3, it is evident that PTLFM outperforms all the other methods and their configurations. Also, it provides better results over the best results of all the competing methods across all the personality traits. The results validate the feature selection using F-statistic, Chi-square, Mutual information with wrapper-based feature selection method. Further, the results with SVM and LR indicate that the method can be used with different classifiers. Moreover, from Table 4 it can be observed that PTLFM provides the best results for the Exact match as well.

It is evident from the Mairesse Baseline's results that the LIWC and MRC combined feature set provides intermediate results and the obtained accuracies are not high enough. The reason might be that it utilizes all the features. The FineEmo, using additional lexicon features, provides some improvements to the Mairesse Baseline's results, and LIWC+FR takes it further using feature selection. The LIWC+FR applies the features selection only on LIWC features using PCA and IG and improves results a little. Hence, these results favor the feature selection approach that the PTLFM method also follows.

The CNN+Mairesse provides comparatively better results than these prevalent methods. The CNN+Mairesse with different configurations is the best-resulted method, where a deep

Table 4 Performance comparison for Exact match

Method	Exact match
Mairesse Baseline with SVM	6.23
FineEmo	6.47
LIWC + FR	6.81
CNN + Mairesse	6.84
PTLFM + LR	7.92
PTLFM + SVM	7.15

learning CNN model with n-gram, i.e., multi-channel architectures with 1-, 2-, 3-, 4-gram, are deployed. Contrary to the preceding best-resulted method, the proposed PTLFM method provisions machine learning and feature selection that utilizes handcrafted features. For handcrafted features, a combination of LIWC and MRC features is used. It is well-known that deep learning methods require a lot more time and space than machine learning methods. Consequently, our method PTLFM improves prediction performance on the one hand and also consumes less time and space on the other hand.

These results suggest that one has to be careful as a popular paradigm is not a remedy for all machine-learning problems. If the number of features and the training set's size is not high, then the traditional approaches would be better or equivalent to deep learning approaches at a lower computational cost. These conclusions align with Occam's razor principle that states to prefer a more straightforward model over complex models among the competing models.

As PTLFM uses a wrapper-based feature selection method, and the time complexity of the wrapper method depends on the classification algorithm used to identify the feature subset and its implementation, the time complexity of PTLFM also depends on the same factor. However, before applying the wrapper approach, PTLFM uses a filter method to reduce the number of features fed to the wrapper method; hence, its complexity will undoubtedly be lower than that of a method that uses all the features, such as sequential feature selection wrapper method.

5.1 Effect of feature selection

This subsection presents the results of experiments performed for feature selection using only one criterion. Instead of one pool as in the earlier experiments, we use three separate lists and a wrapper-based feature selection. The Logistic regression is used as the wrapper classifier. To keep the total number of features in a list and the final number of features the same as in PTLFM, we use $P = 3 * p$. The reason for choosing $P = 3 * p$ is that PTLFM uses three criteria, and the number of features is not greater than $3 * p$ for three lists because some features may be common to two or all three lists.

The experimental results are shown in Table 5, where the best results across a single criterion with a different number of selected features are shown in boldface. The best result across all the criteria for a personality trait are shown as underlined. Here, Chi-only, F-only, and MI-only represent the use of Chi-square, ANOVA's F-statistics, and Mutual information as the single feature selection criteria. Also, we have formed an ensemble of these three methods using a simple voting scheme, i.e., in the ensemble, the final class label is the label favored by two of the three feature selection criteria.

From the Table 5, it is evident that different feature selection criterion with the same number of features provides classification performance differently and the optimal number of selected features are different as well. Further, within a feature selection criteria number of features are different across the personality traits. Comparing the three criteria, i.e., ignoring the ensemble, it is observed that the ANOVA's F-statistic performs best for three personality traits, viz., AGR, CON, and OPN; Mutual information criteria performs best for the remaining two personality traits.Further, ANOVA's F-statistics provides the best average results. However, when the ensemble method is considered, the ensemble method performs slightly better than ANOVA's F-statistic on average and the best on two personality traits. The chi-square is not the best criterion for any personality trait; however, its performance is quite competitive with the other two contrivances. Further, ensemble results suggest that it has helped the ensemble's performance enhancement. This observation also validates

Table 5 Trait wise accuracy for feature selection with a single criteria

Feature	P	p	EXT	NEU	AGR	CON	OPN	Average
	18	6	**57.99**	**60.38**	**58.48**	**58.72**	62.17	**59.55**
	21	7	57.94	60.14	58.25	58.53	62.33	59.44
Chi-only	24	8	57.94	60.34	58.25	58.54	62.28	59.47
	27	9	57.92	60.36	58.18	58.52	62.22	59.44
	30	10	57.91	60.10	57.90	58.46	**62.35**	59.34
	18	6	**58.35**	60.89	58.63	**58.91**	**62.53**	**59.86**
	21	7	58.15	**60.92**	58.50	58.48	62.31	59.67
F-only	24	8	58.06	60.80	58.47	58.49	62.14	59.59
	27	9	58.34	60.66	**58.71**	58.71	62.25	59.73
	30	10	58.26	60.61	58.64	58.65	62.17	59.67
	18	6	58.09	60.83	58.03	58.63	62.44	59.60
	21	7	58.21	61.04	57.96	**58.78**	62.32	59.66
MI-only	24	8	58.27	**61.10**	**58.30**	58.66	62.36	**59.74**
	27	9	58.32	60.90	58.25	58.50	**62.45**	59.68
	30	10	**58.44**	60.76	58.05	58.41	62.32	59.60
	18	6	58.36	60.84	**58.63**	**58.98**	**62.56**	**59.87**
	21	7	58.17	60.98	58.51	58.84	62.53	59.81
Ens	24	8	58.11	**61.03**	58.54	58.80	62.36	59.77
	27	9	58.18	60.92	58.62	58.79	62.38	59.78
	30	10	**58.39**	60.72	58.43	58.64	62.33	59.70

our hypothesis that the three feature selection filter criteria capture different aspects of the training data.

6 Conclusion and future work

This paper presents a method titled *P*ersonality *T*rait classification based on *L*inguistic and *F*eature selection as *M*ulti-label classification (*PTLFM*). The PTLFM method is a feature selection framework based on ranking and selecting high-ranked features using ANOVA's F-statistic, Chi-square, and Mutual information from a combined feature set of LIWC and MRC feature set. The *PTLFM* works in a label-wise paradigm as follows. First, it extracts Linguistic-Inquiry-and-Word Count (LIWC) and Medical Research Council (MRC) features for each label. Then, these features are combined to obtain a single feature space. Next, it selects high ranked features from the combined feature space. To rank features, it uses analysis of variance's F-statistic, Chi-square, and Mutual information. Further, it uses a wrapper-based approach to find the best feature subset from these diverse feature subsets. Finally, it trains a classifier with a balanced weighting scheme. The three feature selection filter methods concentrate on providing the three different perspectives of the dataset. Using a wrapper method PTLFM combines them on the one hand and reduces the number of features on the other hand. The experiment and result analysis ascertain that the proposed method improves all the five personality traits' accuracy and that the combination of three criteria provides better results than the individual criterion.

The linguistic feature construction and feature selection based on F-statistic, Chi-square, Mutual information and wrapper method is a beneficial approach for classification performance improvement. However, this work is only a foundation for identifying the Big-five personality traits more accurately. Further effort is needed, including examining the effect of key feature extraction from other sources than LIWC and MRC, and refining the design of PTLFM accordingly. Also, the ensemble methods may improve classification accuracy using bagging and boosting schemes; however, with increased complexity. Possible future work with ensemble schemes may be considered that combine homogeneous as well as heterogeneous classifiers. As the features from the feature pool are selected randomly, a possible future work may be to induce a heuristic search method instead of random selection. Further, instead of a wrapper-based approach, an evolutionary algorithm may be used in future work.

Data Availability statement This paper reused data and a data citation to the reference list is added in the manuscript.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Aguilar AG, Guillén MJY, Roman NV (2014) Destination brand personality: an application to spanish tourism. Int J Tour Res 18(3):210–219
2. Al Marouf A, Hasan MK, Mahmud H (2020) Comparative analysis of feature selection algorithms for computational personality prediction from social media. IEEE Trans Comput Soc Syst 7(3):587–599
3. Arya R, Singh J, Kumar A (2021) A survey of multidisciplinary domains contributing to affective computing. Comput Sci Rev 40:100399
4. Bergner RM (2020) What is personality? two myths and a definition. New Ideas Psychol 57:100759
5. Bhardwaj S, Atrey PK, Saini MK, El Saddik A (2016) Personality assessment using multiple online social networks. Multimed Tools Appl 75(21):13237–13269
6. Capretz LF, Ahmed F (2010) Making sense of software development and personality types. IT Profession 12(1):6–13
7. Coltheart M (1981) The mrc psycholinguistic database. Quart J Exper Psychol Sect A 33(4):497–505
8. Dhelim S, Aung N, Ning H (2020) Mining user interest based on personality-aware hybrid filtering in social networks. Knowl-Based Syst 206:106227
9. El-Demerdash K, El-Khoribi RA, Shoman MAI, Abdou S (2021) Deep learning based fusion strategies for personality prediction. Egyptian Informatics Journal
10. Elngar AA, Jain N, Sharma D, Negi H, Trehan A, Srivastava A (2020) A deep learning based analysis of the big five personality traits from handwriting samples using image processing. J Inf Technol Manag 12:3–35. Special Issue: Deep Learning for Visual Information Analytics and Management
11. Goldberg LR (1993) The structure of phenotypic personality traits. Am Psychol 48(1):26–34
12. Gulseven O, Mostert J (2019) The role of phenotypic personality traits as dimensions of decision-making styles. Open Psychol J 12(1):84–95
13. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics
14. Lerner MJ, Millon T, Weiner IB (2003) Handbook of psychology, volume 5: personality and social psychology. Wiley
15. Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res 30:457–500
16. Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning-based document modeling for personality detection from text. IEEE Intell Syst 32(2):74–79
17. Mehta Y, Majumder N, Gelbukh A, Cambria E (2019) Recent trends in deep learning based personality detection. Artif Intell Rev:1–27

18. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, vol 26, pp 3111–3119
19. Mishra NK, Singh PK (2020) Fs-mlc: feature selection for multi-label classification using clustering in feature space. Inf Process Manag 57(4):102240
20. Mishra NK, Singh PK (2021) Feature construction and smote-based imbalance handling for multi-label learning. Inf Sci 563:342–357
21. Mishra R, Barnwal SK, Malviya S, Mishra P, Tiwary US (2018) Prosodic feature selection of personality traits for job interview performance. In: International Conference on Intelligent Systems Design and Applications. Springer, pp 673–682
22. Mohammad SM, Kiritchenko S (2015) Using hashtags to capture fine emotion categories from tweets. Comput Intell 31(2):301–326
23. Myers IB (1998) Mbti manual: A guide to the development and use of the myers-briggs type indicator. Consulting Psychologists Press, Palo Alto
24. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. Lawrence Erlbaum Associates, Mahway
25. Pennebaker JW, King LA (1999) Linguistic styles: language use as an individual difference. J Person Soc Psychol 77(6):1296–1312
26. Pohjalainen J, Räsänen O, Kadioglu S (2015) Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. Comput Speech Lang 29(1):145–171
27. Quercia D, Lambiotte R, Stillwell D, Kosinski M, Crowcroft J (2012) The personality of popular facebook users. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp 955–964
28. Rifkin R, Klautau A (2004) In defense of one-vs-all classification. J Mach Learn Res 5:101–141
29. Sharma A, Jayagopi DB (2021) Towards efficient unconstrained handwriting recognition using dilated temporal convolution network. Expert Syst Appl 164:114004
30. Tang B, Kay S, He H (2016) Toward optimal feature selection in naive bayes for text categorization. IEEE Trans Knowl Data Eng 28(9):2508–2521
31. Tayarani M, Esposito A, Vinciarelli A (2019) What an" ehm" leaks about you: Mapping fillers into personality traits with quantum evolutionary feature selection algorithms. IEEE Trans Affect Comput
32. Thakur D, Gera T, Singh J (2015) The senti strength calculator: Engineering the sentiment from the opinionated text. In: 2015 Fifth international conference on communication systems and network technologies. IEEE, pp 1103–1108
33. Tighe EP, Ureta JC, Pollo BAL, Cheng CK, Bulos RDD (2016) Personality trait classification of essays with the application of feature reduction [internet]. In: Proceedings of the 4th workshop on Sentiment Analysis where AI meets Psychology (SAAIP) co-located with 25th International Joint Conference on Artificial Intelligence (IJCAI), pp 22–28
34. Vuttipittayamongkol P, Elyan E (2020) Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Inf Sci 509:47–70
35. Wang C, Han Y (2011) Linking properties of knowledge with innovation performance: the moderate role of absorptive capacity. J Knowl Manag 15(5):802–819
36. Wang Y, Zhao N, Liu X, Karaburun S, Chen M, Zhu T (2020) Identifying big five personality traits through controller area network bus data. J Adv Transp 2020
37. Xue D, Wu L, Hong Z, Guo S, Gao L, Wu Z, Zhong X, Sun J (2018) Deep learning-based personality recognition from text posts of online social networks. Appl Intell 48(11):4232–4246
38. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proceedings of the fourteenth International Conference on Machine Learning, ICML 97. Morgan Kaufmann Publishers Inc., San Francisco, pp 412–420
39. Zhao J, Zeng D, Xiao Y, Che L, Wang M (2020) User personality prediction based on topic preference and sentiment analysis using lstm model. Pattern Recogn Lett 138:397–402
40. Zhao S, Gholaminejad A, Ding G, Gao Y, Han J, Keutzer K (2019) Personalized emotion recognition by personality-aware high-order learning of physiological signals. ACM Trans Multimed Comput Commun Appl 15(1s):1–18

Affiliations

Nitin Kumar Mishra[1] · Aditya Singh[2] · Pramod Kumar Singh[1]

Aditya Singh
aditya.singh2016a@vitalumn.ac.in

Pramod Kumar Singh
pksingh@iiitm.ac.in

[1] Computational Intelligence and Data Mining Research (CIDMR) Lab, ABV-Indian Institute of Information Technology and Management Gwalior, Gwalior 474015, India

[2] Vellore Institute of Technology, Vellore 632014, India