

Predicting personality traits from Arabic text: an investigation of textual and demographic features with feature selection analysis

Khaoula Chraibi, Ilham Chaker, Azeddine Zahi

Department of Computer Science, Faculty of Sciences and Technology, University of Sidi Mohamed Ben Abdellah, Fez, Morocco

Article Info

Article history:

Received Apr 8, 2024

Revised Jul 17, 2024

Accepted Oct 1, 2024

Keywords:

Automatic personality recognition
Big five
Demographic features
Feature selection
Machine learning
Modern standard Arabic

ABSTRACT

Automatic personality recognition (APR) utilizes machine learning to predict personality traits from various data sources. This study aims to predict the big five personality traits from modern standard Arabic (MSA) texts, using both textual and demographic features. The "MSAPersonality" dataset is employed to conduct a comprehensive analysis of features and feature selection methods to evaluate their impact on APR model performance. We compared feature selection algorithms from the filter, wrapper, and embedded-based categories through a systematic experimental design that consisted of feature engineering, feature selection, and regression. This study showed that each trait was more accurately predicted using a distinct set of features. However, age and study level were the most common features among the five traits. Moreover, although there were no statistically significant differences in performance between the feature selection techniques, embedded-based methods offered the best compromise between performance, time, and interpretability. These findings contribute to the understanding of APR in general and among Arabic speakers.

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



Corresponding Author:

Khaoula Chraibi

Faculty of Sciences and Technology, University of Sidi Mohamed Ben Abdellah

B.P. 2202 – Route d'Imouzzer, Fez, Morocco

Email: khaoula.chraibi@usmba.ac.ma

1. INTRODUCTION

Features have a significant impact on the performance and efficiency of machine learning (ML) models. The choice and quality of features directly affect the ability of the model to accurately represent and predict outcomes based on the underlying data patterns. However, including too many features increases data complexity and dimensionality, leading to challenges such as the curse of dimensionality. To overcome these problems, feature selection and dimensionality reduction techniques are often used. They ensure that the model focuses on the most relevant features and reduce the computational complexity of the training process.

The significance of these considerations extends across various ML applications, notably automatic personality recognition (APR). APR stands for the prediction of an individual's self-assessed personality from diverse data sources, including images, textual content, and voice recordings. The development of APR is rooted in the psychological theory of the big five personality traits, which proposes that human personality can be categorized into five major dimensions: openness (OPN), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU) [1].

Empirical research in this area has primarily focused on extracting personality traits from textual data using natural language processing (NLP) techniques. Studies such as those by Pennebaker and King [2],

Yarkoni [3] have demonstrated the feasibility of predicting personality traits from textual content. These studies highlight the potential of textual data to reveal insightful aspects of an individual's personality.

Despite the progress made in this field, the study of personality prediction from modern standard Arabic (MSA) remains underexplored. As a rich and complex language, MSA presents unique challenges and opportunities for APR. The linguistic nuances and cultural specificities inherent in MSA require tailored feature selection and extraction techniques to capture personality traits accurately. This gap is even more pronounced when considering the integration of demographic attributes, such as age and gender, which can provide additional contextual layers to personality prediction models but have yet to be robustly incorporated into research efforts focused on MSA. Furthermore, although some studies have included different feature selection algorithms in the APR domain [4]–[6], comprehensive comparative analyses between various techniques from various categories are scarce. This lack of cross-category comparisons limits our understanding of the relative effectiveness of these algorithms in different contexts and datasets, particularly those involving non-English languages and multimodal data sources.

In light of these considerations, it is imperative to expand the scope of APR research to include underrepresented languages, such as MSA, and holistically integrate demographic factors into personality prediction models. Such efforts not only enrich the APR field but also contribute to the development of more nuanced, culturally sensitive ML applications. To address these gaps, this study used the Arabic personality prediction dataset “MSAPersonality” [7] to investigate two research questions: i) what is the best feature selection method applicable to this APR problem? and ii) Using textual and demographic features, what are the best feature sets for each of the five personality traits?

The remainder of this paper is organized as follows. First, we review several key studies that predict the big five personality traits using textual or demographic data, with particular attention to those that incorporate feature selection into their prediction methodologies. Next, we describe the experimental methodology used in our study. We then present our experimental results, engage in a thorough discussion of these results, and conclude with the key findings and implications of our work.

2. RELATED WORK

In the field of APR, extracting meaningful features from textual and demographic data is crucial for improving the model accuracy and interpretability. This section examines previous research on predicting personality traits using textual and demographic features. In addition, this section explores the various feature selection methods employed in these studies.

2.1. Textual features

Textual features include a wide range of linguistic cues extracted not only from textual input but also from voice and visual inputs. They are invaluable resources for understanding human behavior and personality characteristics. Researchers have increasingly turned to textual analysis to predict the Big Five personality traits of individuals speaking different languages [7].

From the English language, in 2016, researchers extracted lexical features using different dictionaries: Linguistic inquiry and word count (LIWC) in Mukta *et al.* [8], LIWC and Whissell's dictionary of affect in language (DAL) in An *et al.* [9], and LIWC, medical research council Psycholinguistic database (MRC), and NRC (NRC Emotion Lexicon) in Wang *et al.* [10]. Potash *et al.* [11] extracted different linguistic features, including punctuation count, part of speech (POS) count, affin count (frequency of words with an emotional valence), “to” count, and general inquirer tags (the frequency of word tags obtained from the general inquirer tool). In contrast, Celli *et al.* [12] used bag of words (BoW) features.

In 2017, Marwade *et al.* [13] and Iatan [14] used LIWC features, Tandera *et al.* [15] used LIWC and structured programming for linguistic cue extraction (SPLICE) features, whereas [16] used LIWC, Harvard general inquirer (HGI), MRC, and sensorial lexicon (Sensicon). Conversely, Varshney *et al.* [17] used term frequency-inverse document frequency (TF-IDF). In 2018, researchers used LIWC in Zhong *et al.* [18], LIWC and SPLICE in Tadesse *et al.* [19], and LIWC, HGI, MRC, and Sensicon in Kumar *et al.* [20]. Alternatively, Cutler and Kulis [21] used TF-IDF, and Paudel *et al.* [22] used BoW. In 2020, Kafeza *et al.* [23] used LIWC and Moraes *et al.* [24] used TF-IDF. In 2022, Moshkin *et al.* [25] used BoW and TF-IDF.

From a dataset of four languages (English, Spanish, Italian, and Dutch), Arroju *et al.* [26] used TF-IDF and LIWC. Grivas *et al.* [27] used TF-IDF and text statistics (e.g., count of word length, count of capital words). Pervaz *et al.* [28] used a list of stylistic features, such as percentages of different types of punctuation and word types (noun and verb). From Chinese, Peng *et al.* [4] used TF and TF-IDF. Xue *et al.* [29] used linguistic features (such as number of words and punctuation) extracted using a Chinese language psychological analysis system (TextMind), along with social media profile information. Yuan *et al.* [30] also used TextMind. From Indonesian and Arabic, Pratama and Sarno [31], Rumagit and Girsang [32], Huda and Chowanda [33], and Salem *et al.* [34] (Arabic) used TF-IDF.

2.2. Demographic features

Some studies have incorporated demographic information as a feature for predicting personality traits, typically in conjunction with other features rather than standalone characteristics. For example, in 2011, Chapsky [35] used demographic, relational, and cultural attributes as expressed on social media to predict personality. Demographic data consisted of age, sex, educational level, geographic location, and ethnicity. Similarly, Wald *et al.* [36] included demographic information, such as age, gender, location, and relationship status, extracted from social media. Wu and Chen [37] also used gender, age, and education level. Maharjan and Solorio [38] and Ye *et al.* [39] also incorporated age and gender into their feature sets along with other features. In a study conducted by Ding *et al.* [40] in a resort, individuals on vacation were surveyed to collect information about their vacation and various demographic characteristics, including gender, age, income, and number of children.

2.3. Feature selection

Feature selection is an important step in an ML model that focuses on minimizing the dimensionality of the data by retaining the most informative features. While optional, this step has been incorporated in numerous studies. To select the best features from a TF-IDF-based feature set, Peng *et al.* [4] tested the Chi-squared (CHI) test and recursive feature elimination (RFE), whereas Vaidhya *et al.* [41] used principal components analysis (PCA) to reduce the number of features to 10. In contrast, Moreno *et al.* [42] used three different methods: PCA, linear discriminant analysis (LDA), and non-negative matrix factorization. From LIWC-based feature sets or psycholinguistic features in general, Werlen [43] used back-forward propagation feature selection, Tighe *et al.* [44] used both information gain (IG) and PCA, whereas Lin *et al.* [45] used an adapted gray wolf optimizer (GWO).

Regarding other features, Li *et al.* [5] used a two-step approach by first applying correlation analysis to avoid the problem of multiple collinearity, and then PCA to filter and remove the features of low significance. Ding *et al.* [40] used causal feature selection applied on demographic data. Marouf *et al.* [6] compared the performance of five feature selection algorithms: Pearson correlation coefficient, correlation-based feature subset, IG, symmetric uncertainty evaluator, and CHI method with linguistic, psycholinguistic, and social network features. Mishra *et al.* [46] used a method that combines the analysis of variance's *f*-statistic, CHI, and mutual information (MI) with the sequential feature selection wrapper method to rank linguistic features.

3. METHOD

In this section, we detail the methodology used to collect the “MSAPersonality” dataset and the approach used to predict the Big Five personality traits. Our framework, which is analogous to the methodologies employed in the text classification and regression literature [47], comprises four essential steps: data collection, feature engineering, feature selection, and regression. During data collection, personality traits, Arabic writing, and demographic information were collected. Through feature engineering, we prepared the data and extracted meaningful features from the demographic and textual data in the dataset. We then applied feature selection techniques to identify the most relevant features for predicting the Big Five. Finally, we fed the selected features as inputs into various regression models to calculate personality scores for each trait.

3.1. Dataset

“MSAPersonality” was collected through an online survey hosted on a website, from Arabic-speaking participants. The survey was divided into three parts. The first part requested demographic information, such as sex, age range, occupation, level of study, and specialty. The second part of the survey included the Arabic version of the big five inventory (BFI), which was originally proposed by John, Donahue, and Kentle in 1991 [48], and was translated by Al Ansari and Al-Ali in 2018 [49]. The BFI is a 44-item questionnaire that assesses five personality traits using a 5-point Likert scale (5=strongly agree, 1=strongly disagree). In the final part of the survey, participants were asked to write freely about any topic in Arabic with a minimum length requirement of 30 words. To guide their writing, prompts were proposed, such as university experience, work experience, day/week events, and emotions. The survey could not be submitted if any answers were missing, or if a condition was not fulfilled. The survey was distributed through the university's email network and social media groups, and participation was voluntary and unpaid. Participants were informed that their responses would be kept confidential and they received the results of their personality tests after submitting their responses.

Once the data were collected, a manual review was conducted to identify and remove any abnormal text entries. Specifically, responses that contained random letters, text that had been copied and pasted from

the website, text that consisted solely of repeated words, and dialect texts were deleted. Following this initial review, an automated process was used to identify and remove duplicate entries from the dataset. This process ensured that the final dataset contained only valid responses that could be used for further analysis.

The first version of “MSAPersonality” was reported in [7]. The dataset consists of 1,464 entries, with 1,063 female and 401 male respondents. The participants’ ages ranged from 18 to 64 years old. All the participants were either employed (45%) or studying (39%), with the remainder not specified (16%). Regarding educational background, 86% of the participants were higher education students/graduates, 5% had high school degrees, 5% did not complete high school, 1% did not enroll in formal education, and 3% were not specified. Their academic specialties include science, economics, literature, humanities, engineering, rights, medicine, and arts. Each entry in the dataset consisted of demographic data and Arabic text written by the participant, with a minimum length of 30 words. Additionally, each entry included five scores, one for each personality trait measured using the BFI. Trait scores ranged from 1 to 5 and were calculated as the responder’s mean item response (i.e., adding all items scored on a scale and dividing by the number of items on the scale).

3.2. Feature engineering

3.2.1. Demographic features

The dataset includes demographic information, such as gender, age range, occupation, study level, and specialty. To prepare the data for the analysis, we used various encoding techniques tailored to each variable. Binary encoding was used for gender to create a single binary feature representing male/female values. The age range variable is ordinal and was encoded using numerical values corresponding to each age range category. To represent occupation and specialty, we used one-hot encoding to create a binary feature for each unique value within each variable. For the study level, we converted the data to the total number of years of study, enabling us to represent this variable as a continuous numerical feature. By employing these encoding techniques, we transformed the demographic data (DMG) into a format suitable for ML analysis.

3.2.2. Text features

We performed no significant preprocessing of the text. In the feature extraction phase, we extracted linguistic features from Arabic texts using two methods: an LIWC-inspired approach and TF-IDF. LIWC is a widely used text analysis software that analyzes a given text based on a predefined dictionary of words and linguistic categories, providing the frequency of occurrence of each category in the text [50]. Due to the unavailability of an Arabic version, we used a part-of-speech (POS) tagging approach to extract some LIWC linguistic dimensions. After tagging the original text, we calculated the occurrence of different tags equivalent to LIWC dimensions and subdimensions, including pronouns, determiners, prepositions, adverbs, conjunctions, negations, and adjectives. We also included the summary values of the word count, long words, and average words per sentence.

In addition, we used the TF-IDF method to extract features from the text. This method involves calculating the frequency of occurrence of each word in a document, and then scaling the values using the inverse frequency of the word in the entire corpus [51]. This approach helps to identify the most informative words in each document by assigning less weight to commonly occurring words and more weight to rare words that are more specific to each document. The previous steps resulted in the following set of features: {DMG, LIWC, TF-IDF}, which we scaled before proceeding to the next step.

3.3. Feature selection

We employed various feature selection methods to identify the most informative features for personality prediction. These methods belong to three main categories: filter, wrapper, and embedded based feature selection. In the filter-based category, statistical tests are used to evaluate the relevance of each feature and select the most important features. We used mutual information (MI) and univariate feature selection (UFS) using f-regression (FREG). MI quantifies the statistical dependency or information shared between two variables. It measures the amount of information provided by a feature regarding a target variable. The UFS selects the best features based on univariate statistical tests and tests them using f-regression, which calculates the variance f-value between the feature and the target. Feature selection was performed using the SelectKbest algorithm. To determine the value of k (number of features to select), we calculated each feature score using both filter methods and then used the mean, 25th, 50th, and 75th quantiles of the scores as thresholds for selecting features.

In the wrapper-based category, a machine learning model is used to evaluate the performance of various subsets of features and select the best one. For this purpose, we implemented recursive feature elimination (RFE) with cross-validation (RFECV). The RFE operates by sequentially removing the least significant features until a desired number is reached, and cross-validation is used to determine this number.

In other words, RFECV evaluates various subsets of features and selects the subset with the highest score. We applied the wrapper method in conjunction with a random forest regressor.

For the embedded-based category, feature selection is combined with the training of a machine learning model. We used two methods: random forest and LASSO. Random forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy. It can also provide information on feature importance. LASSO is a linear regression model that can perform feature selection by lowering the coefficients of less important features to zero. The selection of features was performed using the `SelectFromModel` algorithm.

3.4. Regression

To predict personality traits based on the selected features, we employed various regression methods representing different regression categories. Specifically, we used the following models: random forest (RF), k-nearest neighbors (KNN), support vector regressor (SVR), linear regression (LR), and Bayesian ridge (BR). The RF is an ensemble method that constructs multiple decision trees and averages their outputs to obtain accurate predictions. KNN computes the target value for a data point by averaging the values of its k-nearest neighbors. SVR uses support vectors to define a hyperplane that maximizes the margin between the target variable and the predicted values. LR fits a linear model to the data by minimizing the sum of the squared errors. The BR estimates the parameters of a linear model using Bayesian inference.

Using regression methods that represent different categories of regression techniques (tree-based, instance-based, kernel-based, linear-based, and probabilistic-based methods), we aimed to explore the potential of each method and provide a robust analysis of the data. The performance of these models was evaluated using the root mean squared error (RMSE). The RMSE measures the average deviation of the predicted values from the actual values and is calculated as the square root of the average squared difference between the predicted and actual values.

4. RESULTS

4.1. Experimental setup

The implementations were performed using Python in the Jupyter lab [52]. We used the CAMEL Tools for POS tagging [53] and scikit-learn for machine learning functions [54]. The calculations were run on a machine with a Linux Ubuntu 22.04.4 LTS operating system and 16 Intel(R) Xeon(R) CPU E5-2637 v4 @ 3.50 GHz processors with 32 GB of RAM. During model development, the data were split into training (70%) and test (30%) sets.

4.2. Experimental results

The experimentations in this study were divided into three distinct experiments with five different scenarios in total, each using a different set of features and regression methods.

- a) Experiment 1 (using features separately): In this experiment, we conducted three scenarios, each using only one feature category. Specifically, scenario 1 (S1) employed LIWC features, scenario 2 (S2) used DMG features, and scenario 3 (S3) used TF-IDF features. The results for each personality trait and outcomes obtained from the five regression models are presented in Table 1. For S1 and S2, we achieved the best results with BR for all traits, except for NEU, which achieved the best result with LR. In S3, the best results were obtained using RF and SVR. In contrast, AGR, CON, and EXT had the best results in S2, whereas NEU and OPN achieved the best results in S3.
- b) Experiment 2 (feature categories combination): In this experiment, we combined features in two distinct scenarios. Scenario 4 (S4) combined the LIWC and DMG features, whereas scenario 5 (S5) combined the LIWC, DMG, and TF-IDF features. The results are presented in Table 2. In S4, the best results were obtained with BR for all traits, except for NEU, which had the best results with RF. Conversely, in S5, the best results were achieved with RF for all traits except AGR, which performed better with SVR. Notably, AGR and EXT achieved the best results in S4, whereas CON, NEU, and OPN achieved the best performance in S5.
- c) Experiment 3 (introducing feature selection): We included the feature selection phase in scenario S5, which combines all features. After implementing Feature Selection, the LR results noticeably deteriorated to the extent that they became inconclusive; hence, we opted not to include them in this experiment. The average execution time (training + testing) for each feature selection method and the standard deviation are also reported. As shown in Table 3, the RF consistently outperformed the other regression models in most cases, yielding the best results for all personality traits. Regarding feature selection techniques, we observed that NEU and OPN had the best results with RFECV, whereas AGR, CON, and EXT

demonstrated the highest performances with FREG, LASSO, and MI, respectively. Table 4 summarizes the best results for all traits across the five scenarios.

Table 1. Experiment 1 results

Scenario	Trait	Regressor	BR	KNN	LR	RF	SVR
S1: LIWC Features (#Features: 26)	AGR		0.486703	0.532876	0.495192	0.501479	0.510986
	CON		0.606752	0.664381	0.617135	0.614387	0.641593
	EXT		0.667577	0.711406	0.678323	0.684746	0.703471
	NEU		0.779079	0.850219	0.777055	0.785002	0.793368
	OPN		0.503429	0.540369	0.509988	0.507629	0.535681
S2: Demographic Features (#Features: 15)	AGR		0.479033	0.515335	0.480160	0.521580	0.488703
	CON		0.590585	0.635165	0.591678	0.666381	0.596017
	EXT		0.654443	0.678375	0.656286	0.696813	0.667148
	NEU		0.774264	0.842234	0.773539	0.898875	0.784582
	OPN		0.507092	0.565864	0.507540	0.580315	0.521853
S3: Text Features - TF-IDF (#Features: 16916)	AGR		0.484387	0.487191	5.378336e10	0.480270	0.483800
	CON		0.610845	0.699836	1.074423e11	0.594230	0.609722
	EXT		0.674662	0.696131	0.674515	0.674105	0.666615
	NEU		0.802922	0.885798	1.806985e11	0.765049	0.800253
	OPN		0.501951	0.657873	9.767487e10	0.507054	0.498602

Bold values indicate the best RMSE for a specific scenario and trait, whereas italic values indicate the best RMSE for the trait across all scenarios.

Table 2. Experiment 2 results

Scenario	Trait	Regressor	BR	KNN	LR	RF	SVR
S4: LIWC + Demographic (#Features: 57)	AGR		0.482665	0.510922	0.488230	0.488991	0.498038
	CON		0.586843	0.644681	0.592422	0.597222	0.623602
	EXT		0.659279	0.722476	0.667086	0.663476	0.689160
	NEU		0.763400	0.849920	0.758073	0.756245	0.777204
	OPN		0.503322	0.544950	0.510910	0.504950	0.532060
S5: LIWC + Demographic + TF-IDF (#Features: 16973)	AGR		0.484123	0.486686	5.834797e10	0.483470	0.483431
	CON		0.605359	0.702062	1.166959e11	0.581489	0.606659
	EXT		0.673811	0.696380	0.674635	0.663497	0.666002
	NEU		0.799886	0.886390	1.962613e11	0.748698	0.798364
	OPN		0.501193	0.657349	1.061964e11	0.496889	0.498183

Bold values indicate the best RMSE for a specific scenario and trait, whereas italic values indicate the best RMSE for the trait across all scenarios.

Table 3. Experiment 3 results

Scenario + FS method	#Features	Avg Time(std)	Regressor	BR	KNN	RF	SVR
S5 + FREG	12718	5,3 (5,1)s	AGR	0.494131	0.486056	0.480596	0.481225
	4240	5,2 (4,3)s	CON	0.604500	0.623472	0.582960	0.606763
	12718	5,5 (5,4)s	EXT	0.677113	0.665809	0.662817	0.666380
	12906	13,7 (12,3)s	NEU	0.797602	0.812971	0.751747	0.798087
	12955	11,5 (12,6)s	OPN	0.502298	0.504666	0.498653	0.498941
S5 + MI	5780	82,1 (4,4)s	AGR	0.483428	0.518336	0.485449	0.484545
	5771	81,6 (3,4)s	CON	0.608455	0.633731	0.580762	0.605302
	5834	81,6 (5,2)s	EXT	0.704719	0.682072	0.657133	0.665668
	5722	81,2 (3,8)s	NEU	0.795188	0.882078	0.758346	0.790967
	5826	83,8 (5,6)s	OPN	0.504325	0.515637	0.510813	0.503671
S5 + RFECV	7957	67812,0 (12,4)s	AGR	0.485477	0.497938	0.483293	0.482498
	11137	41410,6 (10,5)s	CON	0.603617	0.638559	0.582567	0.609830
	7497	73294,6 (12,9)s	EXT	0.701716	0.738238	0.663947	0.668286
	13617	37712,2 (14,0)s	NEU	0.793199	0.875255	0.750442	0.797263
	1337	86115,8 (4,8)s	OPN	0.549400	0.593545	0.497184	0.505406
S5 + LASSO	1785	18,9 (2,7)s	AGR	0.586021	0.498411	0.491473	0.495228
	1594	16,4 (1,5)s	CON	0.695677	0.628539	0.573859	0.616613
	1596	19,1 (2,8)s	EXT	0.801470	0.681394	0.671761	0.695437
	1623	18,5 (1,7)s	NEU	0.894125	0.817067	0.761368	0.816641
	1506	19,1 (3,1)s	OPN	0.612719	0.515323	0.518997	0.525137
S5 + RF	1075	57,8 (3,8)s	AGR	0.526346	0.499595	0.481272	0.507034
	942	38,9 (2,5)s	CON	0.666515	0.622796	0.579713	0.608439
	1077	63,1 (4,1)s	EXT	0.734034	0.674910	0.662721	0.678823
	931	42,7 (2,7)s	NEU	0.835743	0.793378	0.751364	0.791816
	1090	65,1 (4,2)s	OPN	0.550221	0.603719	0.498040	0.511702

Bold values indicate the best RMSE for a specific scenario and trait, whereas italic values indicate the best RMSE for the trait across all scenarios.

Table 4. Results summary

Scenario	Trait	AGR	CON	EXT	NEU	OPN
	FS method					
S1: LIWC features	none	0.486703	0.606752	0.667577	0.777055	0.503429
S2: Demographic features	none	0.479033	0.590585	0.654443	0.773539	0.507092
S3: Text features - TF-IDF	none	0.480270	0.594230	0.666615	0.765049	0.498602
S4: LIWC + Demographic	none	0.482665	0.586843	0.659279	0.756245	0.503322
S5: LIWC + Demographic + TF-IDF	none	0.483431	0.581489	0.663497	0.748698	0.496889
	FREG	0.480596	0.582960	0.662817	0.751747	0.498653
	MI	0.483428	0.580762	0.657133	0.758346	0.503671
	RFECV	0.482498	0.582567	0.663947	0.750442	0.497184
	LASSO	0.491473	0.573859	0.671761	0.761368	0.515323
	RF	0.481272	0.579713	0.662721	0.751364	0.498040

5. DISCUSSION

5.1. Insights into feature selection techniques

We presented the outcomes of five feature selection techniques, which are classified into three categories: Filter, Wrapper, and Embedded methods, as shown in Table 3. Our findings revealed that RFECV achieved the best results for NEU and OPN, whereas FREG, LASSO, and MI demonstrated the best performance for AGR, CON, and EXT, respectively. This indicates that the optimal technique varies, depending on the target trait. Despite these variations, an independent samples t-test indicated no statistically significant differences between the methods (p -values $> \alpha=0.05$). This implies that, based on performance alone, it is not possible to select a single technique as superior to the others.

Additionally, in terms of execution time, FREG, LASSO, RF, and MI had relatively small differences, while RFECV took significantly longer. This longer duration, even with better performance, poses a challenge for model enhancements and retraining, a common drawback of wrapper methods [55]. Regarding the number of features used, the reduced feature set from RF and LASSO not only simplifies the model, but also enhances interpretability, a critical aspect in ML.

To the best of our knowledge, no study has compared all three feature selection categories in the context of APR. However, some studies compared two categories or methods within a single category. For example, [4] compared two categories and found that CHI (Filter) performed better than RFE (Wrapper), which is contrary to our findings. Within categories, [40] compared two wrapper methods and found that fast greedy equivalence search outperformed the PC algorithm. Among filter-based methods, [6] found that the Pearson correlation coefficient was the best performer, surpassing the correlation-based feature subset, IG, symmetric uncertainty evaluator, and CHI methods.

In conclusion, when considering the performance, time, and interpretability, embedded feature selection methods (such as RF and LASSO) are the most suitable approaches for this problem. To strengthen our conclusion and extend its applicability to the MSA-based APR problem, further testing on additional datasets, when available, and evaluation of feature selection methods with various features are recommended.

5.2. Insights into the best selected features

Upon examining the best results across all scenarios, as shown in Table 4, it is apparent that two scenarios surpassed the others. Both AGR and EXT achieved the highest results with S2, whereas CON, NEU, and OPN demonstrated the best outcomes with S5. The optimal sets of features that yielded the best results for each trait are determined as follow:

AGR={DMG}

CON={DMG(age, studylevel); LIWC*(‘QMark’, ‘i’, ‘adverb’, ‘negate’); TF-IDF(1588 words)}

EXT={DMG}

NEU={DMG; LIWC; TF-IDF}

OPN={DMG; LIWC; TF-IDF}

Union=AGR \cup CON \cup EXT \cup NEU \cup OPN={DMG; LIWC; TF-IDF}

Intersection=AGR \cap CON \cap EXT \cap NEU \cap OPN={DMG (age, studylevel)}

(*The LIWC features in CON stands for the percentage of: question marks ‘QMark’, 1st person singular ‘i’, adverbs ‘adverb’, negations ‘negate’)

It is evident from the aforementioned set of features, their union, and intersection that each feature plays a role in predicting a specific personality trait. However, it is worth noting that age and study level features are common across all five traits. In a study conducted by Ding *et al.* [40], age was found to be predictive of AGR, NEU, and OPN, while gender was predictive of AGR and EXT. Furthermore, Wu and Chen [37] also utilized gender, age, and education level in their study and found that age was correlated with OPN. Therefore, it can be concluded that these demographic factors are significant predictors

of personality traits. In addition, Wald *et al.* [36] and Maharjan *et al.* [38] predicted the five personality traits using a combination of demographic and textual features.

Our findings contribute to the existing literature by demonstrating and confirming that each personality trait can have a distinct set of predictors [6], even within the Arabic context. Notably, demographic features, which are not linguistically dependent, have shown consistent predictive power and can be generalized across different contexts.

6. CONCLUSION

As the field of APR continues to evolve, it is important to understand how personality can be predicted for non-English speakers. In this study, we used the MSA Arabic dataset “MSAPersonality” to explore the impact of different features and feature selection methods on the performance of the APR regression problem. We extracted LIWC-based features, TF-IDF, and demographic features and tested their predictive ability for personality across different scenarios and with different regressors. Moreover, we applied five distinct feature selection methods to reduce the dimensionality of a combined set of all features. Although the improvement in performance was not statistically significant, the impact on the model’s interpretability and run time is important, as a smaller number of features is more understandable and faster to execute than a larger number. We deduced that embedded-based methods offer the best compromise between performance, time, and interpretability. Each trait could be predicted more accurately using a distinct set of features, although age and study level were the most common features among the five traits. The results of our study provide a foundation for further research on personality in an Arabic context and for advancing the Arabic APR. Future studies could investigate our framework outcomes on other datasets, the integration of more advanced NLP techniques for feature extraction, and the use of deep learning models for personality prediction.

REFERENCES




- [1] J. Votano and M. Parham, “The big five trait taxonomy: History, measurement, and theoretical perspectives,” *Handbook of personality: Theory and research*, no. 510, 1999.
- [2] 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.
- [3] T. Yarkoni, “Personality in 100,000 Words: a large-scale analysis of personality and word use among bloggers,” *Journal of Research in Personality*, vol. 44, no. 3, pp. 363–373, Jun. 2010, doi: 10.1016/j.jrp.2010.04.001.
- [4] K.-H. Peng, L.-H. Liou, C.-S. Chang, and D.-S. Lee, “Predicting personality traits of Chinese users based on Facebook wall posts,” in *2015 24th Wireless and Optical Communication Conference (WOCC)*, Oct. 2015, pp. 9–14, doi: 10.1109/WOCC.2015.7346106.
- [5] C. Li, J. Wan, and B. Wang, “Personality prediction of social network users,” in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*, Oct. 2017, pp. 84–87, doi: 10.1109/DCABES.2017.25.
- [6] A. Al Marouf, M. K. Hasan, and H. Mahmud, “Comparative analysis of feature selection algorithms for computational personality prediction from social media,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 587–599, Jun. 2020, doi: 10.1109/TCSS.2020.2966910.
- [7] K. Chraïbi, I. Chaker, Y. Dhassi, and A. Zahi, “MSAPersonality: a modern standard Arabic dataset for personality recognition,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 4, pp. 4498–4507, Aug. 2024, doi: 10.11591/ijece.v14i4.pp4498-4507.
- [8] M. S. H. Mukta, M. E. Ali, and J. Mahmud, “Identifying and validating personality traits-based homophilies for an egocentric network,” *Social Network Analysis and Mining*, vol. 6, no. 1, Dec. 2016, doi: 10.1007/s13278-016-0383-4.
- [9] G. An, S. I. Levitan, R. Levitan, A. Rosenberg, M. Levine, and J. Hirschberg, “Automatically classifying self-rated personality scores from speech,” in *Interspeech 2016*, Sep. 2016, pp. 1412–1416, doi: 10.21437/Interspeech.2016-1328.
- [10] S. Wang *et al.*, “VirtualIdentity: privacy preserving user profiling,” in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2016, pp. 1434–1437, doi: 10.1109/ASONAM.2016.7752438.
- [11] P. Potash and A. Rumshisky, “Recommender system incorporating user personality profile through analysis of written reviews,” *CEUR Workshop Proceedings*, vol. 1680, pp. 60–66, 2016.
- [12] F. Celli, A. Ghosh, F. Alam, and G. Riccardi, “In the mood for sharing contents: Emotions, personality and interaction styles in the diffusion of news,” *Information Processing and Management*, vol. 52, no. 1, pp. 93–98, Jan. 2016, doi: 10.1016/j.ipm.2015.08.002.
- [13] A. Marwade, N. Kumar, S. Mundada, and J. Aghav, “Augmenting e-commerce product recommendations by analyzing customer personality,” in *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*, Sep. 2017, pp. 174–180, doi: 10.1109/CICN.2017.8319380.
- [14] I. F. Iatan, “Predicting human personality from social media using a fuzzy neural network,” *Studies in Computational Intelligence*, vol. 661, pp. 81–105, 2017, doi: 10.1007/978-3-319-43871-9_3.
- [15] T. Tandra, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetyo, “Personality prediction system from Facebook users,” *Procedia Computer Science*, vol. 116, pp. 604–611, 2017, doi: 10.1016/j.procs.2017.10.016.
- [16] T. Maheshwari, A. N. Reganti, T. Chakraborty, and A. Das, “Socio-ethnic ingredients of social network communities,” in *CSCW 2017 - Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 235–238, doi: 10.1145/3022198.3026322.
- [17] V. Varshney, A. Varshney, T. Ahmad, and A. M. Khan, “Recognising personality traits using social media,” in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Sep. 2017, pp. 2876–2881, doi:

- 10.1109/ICPCSI.2017.8392248.
- [18] X. F. Zhong, S. Z. Guo, L. Gao, H. Shan, and D. Xue, "A general personality prediction framework based on Facebook profiles," in *ACM International Conference Proceeding Series*, 2018, pp. 269–275, doi: 10.1145/3195106.3195124.
 - [19] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality predictions based on user behavior on the Facebook social media platform," *IEEE Access*, vol. 6, pp. 61959–61969, 2018, doi: 10.1109/ACCESS.2018.2876502.
 - [20] U. Kumar, A. N. Reganti, T. Maheshwari, T. Chakroborty, B. Gambäck, and A. Das, "Inducing personalities and values from language use in social network communities," *Information Systems Frontiers*, vol. 20, no. 6, pp. 1219–1240, Dec. 2018, doi: 10.1007/s10796-017-9793-8.
 - [21] A. Cutler and B. Kulis, "Inferring human traits from facebook statuses," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11185 LNCS, pp. 167–195, 2018, doi: 10.1007/978-3-030-01129-1_11.
 - [22] A. Paudel, B. R. Bajracharya, M. Ghimire, N. Bhattarai, and D. S. Baral, "Using personality traits information from social media for music recommendation," in *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, Oct. 2018, pp. 116–121, doi: 10.1109/CCCS.2018.8586831.
 - [23] E. Kafeza, A. Kanavos, C. Makris, G. Pispirigos, and P. Vikatos, "T-PCCE: Twitter personality based communicative communities extraction system for big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1625–1638, Aug. 2020, doi: 10.1109/TKDE.2019.2906197.
 - [24] R. Moraes, L. L. Pinto, M. Pilankar, and P. Rane, "Personality assessment using social media for hiring candidates," in *2020 3rd International Conference on Communication Systems, Computing and IT Applications, CSCITA 2020 - Proceedings*, 2020, pp. 192–197, doi: 10.1109/CSCITA47329.2020.9137818.
 - [25] V. Moshkin, N. Yarushkina, and R. Shakurov, "An approach to the psycholinguistic analysis of social media texts using the big five personality traits," in *Lecture Notes in Networks and Systems*, 2022, pp. 479–488.
 - [26] M. Arroju, A. Hassan, and G. Farnadi, "Age, gender and personality recognition using Tweets in a multilingual setting: notebook for PAN at CLEF 2015," in *Conference and Labs of the Evaluation Forum*, 2015.
 - [27] A. Grivas, A. Krithara, and G. Giannakopoulos, "Author profiling using stylometric and structural feature groupings," *CEUR Workshop Proceedings*, vol. 1391, 2015.
 - [28] I. Pervaz, I. Ameer, A. Sittar, R. Muhammad, and A. Nawab, "Identification of author personality traits using stylistic features: notebook for PAN at CLEF 2015," *CLEF 2015 Labs and Workshops, Notebook Papers*, 2015.
 - [29] D. Xue *et al.*, "Personality recognition on social media with label distribution learning," *IEEE Access*, vol. 5, pp. 13478–13488, 2017, doi: 10.1109/ACCESS.2017.2719018.
 - [30] C. Yuan, Y. Hong, and J. Wu, "Personality expression and recognition in Chinese language usage," *User Modeling and User-Adapted Interaction*, vol. 31, no. 1, pp. 121–147, Mar. 2021, doi: 10.1007/s11257-020-09276-2.
 - [31] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," in *2015 International Conference on Data and Software Engineering (ICoDSE)*, Nov. 2015, pp. 170–174, doi: 10.1109/ICoDSE.2015.7436992.
 - [32] R. Y. Rumagit and A. S. Girsang, "Predicting personality traits of facebook users using text mining," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 20, 2018.
 - [33] S. Huda and A. Chowanda, "Personality prediction from text on social media with machine learning," *ICIC Express Letters*, vol. 15, no. 12, pp. 1243–1251, 2021.
 - [34] M. S. Salem, S. S. Ismail, and M. Aref, "Personality traits for egyptian twitter users dataset," in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, Apr. 2019, pp. 206–211, doi: 10.1145/3328833.3328851.
 - [35] D. Chapsky, "Leveraging online social networks and external data sources to predict personality," in *2011 International Conference on Advances in Social Networks Analysis and Mining*, Jul. 2011, pp. 428–433, doi: 10.1109/ASONAM.2011.121.
 - [36] R. Wald, T. Khoshgoftaar, and C. Sumner, "Machine prediction of personality from Facebook profiles," in *2012 IEEE 13th International Conference on Information Reuse and Integration (IRI)*, Aug. 2012, pp. 109–115, doi: 10.1109/IRI.2012.6302998.
 - [37] W. Wu and L. Chen, "Implicit acquisition of user personality for augmenting movie recommendations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, pp. 302–314.
 - [38] S. Maharjan and T. Solorio, "Using wide range of features for author profiling," *CEUR Workshop Proceedings*, vol. 1391, 2015.
 - [39] Z. Ye, Y. Du, and L. Zhao, "Predicting personality traits of users in social networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10585 LNCS, pp. 181–191, 2017, doi: 10.1007/978-3-319-68935-7_21.
 - [40] T. Ding, C. Zhang, and M. Bos, "Causal feature selection for individual characteristics prediction," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2018, pp. 540–547, doi: 10.1109/ICTAI.2018.00089.
 - [41] M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, and A. Shakya, "Personality traits analysis from Facebook data," in *2017 21st International Computer Science and Engineering Conference (ICSEC)*, Nov. 2017, pp. 1–5, doi: 10.1109/ICSEC.2017.8443932.
 - [42] D. R. Jaimes Moreno, J. Carlos Gomez, D.-L. Almanza-Ojeda, and M.-A. Ibarra-Manzano, "Prediction of personality traits in twitter users with latent features," in *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Feb. 2019, pp. 176–181, doi: 10.1109/CONIELECOMP.2019.8673242.
 - [43] L. M. Werlen, "Statistical learning methods for profiling analysis," *CEUR Workshop Proceedings*, vol. 1391, 2015.
 - [44] E. P. Tighe, J. C. Ureta, B. A. L. Pollo, C. K. Cheng, and R. De Dios Bulos, "Personality trait classification of essays with the application of feature reduction," *CEUR Workshop Proceedings*, vol. 1619, pp. 22–28, 2016.
 - [45] H. Lin, C. Wang, and Q. Hao, "A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed Gray Wolf Optimizer for feature selection," *Information Processing and Management*, vol. 60, no. 2, Mar. 2023, doi: 10.1016/j.ipm.2022.103217.
 - [46] 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, Jun. 2022, doi: 10.1007/s11042-022-12548-1.
 - [47] V. Dogra *et al.*, "A complete process of text classification system using state-of-the-art NLP models," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–26, Jun. 2022, doi: 10.1155/2022/1883698.
 - [48] B. Rammstedt and O. P. John, "Big five inventory," *Encyclopedia of Personality and Individual Differences*, pp. 1–4, 2017, doi: 10.1007/978-3-319-28099-8_445-1.
 - [49] T. B. AlAli and B. M. Al Ansari, "Psychometric properties of the Arabic version of big five inventory in a sample of university students in Kuwait," *Journal of Educational and Psychological Sciences*, vol. 19, no. 02, pp. 167–203, Jun. 2018, doi: 10.12785/JEPS/190206.




- [50] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010, doi: 10.1177/0261927X09351676.
- [51] A. Zheng and A. Casari, "Feature engineering for machine learning: principles and techniques for data scientists," *O'Reilly*, 2018.
- [52] T. Kluyver *et al.*, "Jupyter Notebooks—a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*, 2016, pp. 87–90, doi: 10.3233/978-1-61499-649-1-87.
- [53] O. Obeid *et al.*, "CAMEL tools: An open source python toolkit for arabic natural language processing," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020*, pp. 7022–7032.
- [54] F. Pedregosa *et al.*, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [55] N. Nnamoko, F. Arshad, D. England, J. Vora, and J. Norman, "Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning," in *The 15th Annual Postgraduate Symposium on the convergence of Telecommunication, Networking and Broadcasting*, 2014, no. JUNE, pp. 63–67.

BIOGRAPHIES OF AUTHORS






Khaoula Chraibi    received her master's degree in intelligent systems and networks from the Faculty of Sciences and Technology FST, in the University of Sidi Mohamed Ben Abdellah USMBA, Fez, Morocco, in 2018. She is currently a Ph.D. student at the intelligent systems and applications laboratory at FST, Fez. Her research interests include machine learning, affective computing, and personality prediction. She can be contacted at email: khaoula.chraibi@usmba.ac.ma.



Ilham Chaker    received a Ph.D. degree in computer science from the University of Sidi Mohamed Ben Abdellah USMBA, Fez, in 2011. She is a Professor of Computer Science in the Faculty of Sciences and Technology, USMBA Fez, and a member of the Laboratory of Intelligent Systems and Applications. Her main research areas include machine learning, optical character recognition, knowledge management, and human-computer interaction. She can be contacted at email: ilham.chaker@usmba.ac.ma.



Azeddine Zahi    received his Ph.D. degree in 1997 in computer sciences from Mohammed V University in Rabat. He is currently a research professor at Sidi Mohamed Ben Abdellah University of Fez since 1995. His research interests include data science, data mining, artificial intelligence, and ad hoc network. He can be contacted at email: azeddine.zahi@usmba.ac.ma.