MSAPersonality: a modern standard Arabic dataset for personality recognition

Khaoula Chraibi, Ilham Chaker, Younes Dhassi, Azeddine Zahi
Department of Computer Science, Faculty of Sciences and Technology, University of Sidi Mohamed Ben Abdellah, Fez, Morocco

Article Info

Article history:

Received Jul 19, 2023
Revised Feb 13, 2024
Accepted Mar 5, 2024

Keywords:

Affective computing
Arabic text analysis
Automatic personality recognition
Machine learning
Natural language processing

ABSTRACT

Automatic personality recognition is a task that attempts to automatically infer personality traits from a variety of data sources, including Text. Our words, whether spoken or written, reveal a lot about who we are. As people speak different languages, each with its own set of characteristics and level of complexity, identifying their personalities automatically might be language-dependent. This task requires an annotated text corpus with personality traits. However, the lack of corpora for languages other than English makes the task extremely challenging. We concentrated our efforts in this paper on the Arabic language in particular because it is understudied and lacks a corpus, despite being one of the most widely spoken languages in the world. Our primary goal was constructing our "MSAPersonality" dataset, which consists of 267 texts in modern standard Arabic that have been annotated with the Big Five personality traits. To evaluate the dataset and its potential for classification and regression, we used text preprocessing techniques, feature extraction, and machine learning algorithms. We obtained promising experimental results. Therefore, further research into predicting personality from Arabic text can be conducted.

Corresponding Author:

Khaoula Chraibi
Department of Computer Science, 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

People's decisions and preferences are influenced by their personalities. This information is useful in various fields including website/social media recommendations, marketing, advertising, education, and healthcare. Personality can be described according to various models, including the Myers-Briggs type indicator (MBTI), 16 personality factors, and the Big Five. Thus far, the latter model is the most widely recognized paradigm in psychology [1]. It was created by Costa and McCrae [2], and organizes all personality traits into five essential factors: openness (OPN), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU).

Personality can be determined using direct methods. This entails a person taking psychological tests, either with or without the assistance of an expert or having an outsider assess their personality. Because this method is time consuming and requires individual intervention, it can be difficult to implement on a large scale. On the other hand, people leave a trail of traces and data, both digital and non-digital, that can be used to infer their personality indirectly. For example, language use, whether spoken or written, is a strong indicator [3]–[5]. The automation of this process has been the main focus of the personality computing field in computer science [6]. One of the problems addressed is automatic personality recognition (APR). APR stands for self-assessed personality prediction, as opposed to automatic personality perception, which predicts personality based on how others perceive it.

Personality can be predicted from digital data such as images, text, voice, video, and digital behavior. In 2019, we performed a systematic mapping study (SMS) on automatic personality prediction (APPR) that covered the period between 2003 and mid-2019 [7]. Out of 379 papers published on this topic, 122 used texts only, and 36 combined texts with other data (such as voice and social media profiles).

Text data are language dependent. From Figure 1, we can see that 17 different languages have been studied, with English being the most studied. Spanish, Dutch, Italian, Chinese, and Indonesian languages have been researched less frequently. Others, such as Arabic, barely appeared in one or two papers.

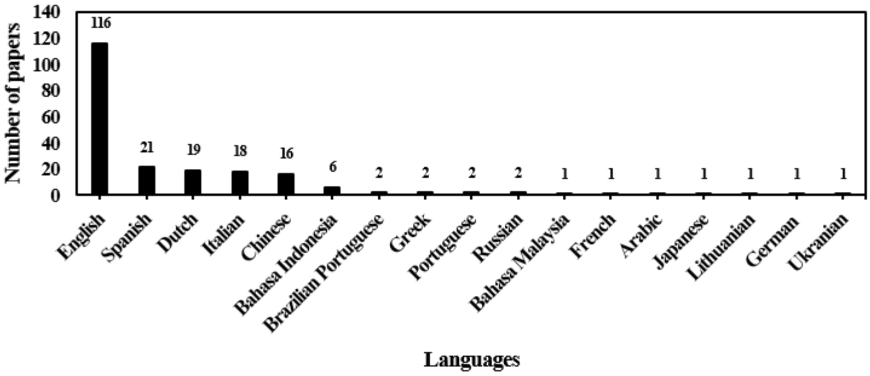

Figure 1. Number of papers per language

The Arabic language is spoken by 315 million native speakers [8] and even more as a second language. There are three primary types of Arabic. Religious texts and many old Arabic manuscripts use classical Arabic, commonly known as Quranic Arabic. The formal language of communication understood by the majority of Arabic speakers is modern standard Arabic (MSA). In everyday communication, Arabic dialects specific to each country or region are used.

Arabic is written and read from right to left. It has a 28-letter alphabet, with no upper or lower cases. It employs diacritical marks, which can be placed above or below letters, to correctly pronounce a word and clarify its meaning. Diacritical marks, on the other hand, are not required for fluent speakers to comprehend a given text.

Furthermore, a language-specific dataset is required to predict personality. Corpora such as myPersonality [9], PAN-AP-2015 [10], and Essays [3] are frequently used in literature [7]. David Stillwell and Michal Kosinski collected the myPersonality dataset via the myPersonality Facebook app from 2007-2012. The application provided Facebook users with a set of genuine personality and ability tests, including a Big Five personality test. The users were then asked for permission to record the information stored on their Facebook profiles anonymously. Approximately 40% of over 6 million users agreed. The owners shared part of the anonymized data with other scholars for non-commercial academic research. However, this dataset is no longer available as of 2018. Several studies have used this dataset [11], [12].

The PAN-AP-2015 corpus was created for the author profiling shared task organized at PAN 2015. This task aimed to identify the age, gender, and personality traits of Twitter users. Data were collected from Twitter in English, Spanish, Italian, and Dutch. The corpus was annotated with self-assessed Big Five personality traits using a short online test. It provides 336, 228, 86, and 76 users' information for English, Spanish, Italian, and Dutch, respectively. The dataset has been a subject of investigation in studies such as [13]–[15], among others.

The Essays dataset was created between 1992 and 2004 [3]. Students in introductory psychology classes were given two assignments to complete each at the beginning and end of the semester. For the first, they were instructed to write in a stream-of-consciousness style, allowing them to freely express their thoughts and feelings. In the second task, they were asked to convey their feelings and thoughts about coming to or being in college. The essays were submitted electronically or on diskettes. The students were provided with a series of questionnaires to fill out, including a Big Five test. Those who turned in at least two assignments with a minimum of 30 words were included in the analysis. The resulting corpus contains 2,468 English essays labeled with self-assessed personality traits, and it has been explored in several studies, such as [16]–[18], to name a few.

Some researchers have created datasets to suit their needs and case studies. Their process is not far from that of the previously described corpora. Data can be gathered in either a closed or open setting. The closed setting entails inviting a specific and well-defined group of people to participate, such as college or university students [19]–[22], graduate course attendees [23], or a group of friends [24]. In an open setting, people from different environments are encouraged to volunteer primarily through social media [25]–[30] and email [31], [32]. Alternatively, accessible online data can be collected without the involvement of any participant [33], [34].

To acquire text data, researchers can collect existing text from social media (Facebook [20], [24], [29], [30], [32], Twitter [27], [35]–[38], Sina Weibo [25], [26], [28], and others [21]), web blogs [39], and email [31]. They could also ask participants to write free text about anything they wanted [19], [29], [40], or controlled text about a particular topic [23], [29]. The written text can be submitted directly in text format or handed to experimenters on paper for later typing [40].

To obtain ground-truth labels for personality traits, participants are typically asked to complete a personality test [19], [21], [23]–[32], [36], [38]–[40]. Otherwise, personality labels can be obtained by searching for people sharing this information on social media [34] or by labeling the data using well-established personality prediction methods [33], [41]. Finally, personality can be predicted using a lexicon-based approach, eliminating the need for labels [42].

The papers cited above studied different languages. Two papers were found to be particularly interested in Arabic. Alsadhan and Skillicorn [34] proposed a language-independent approach that they applied to eight languages including Arabic. They collected an Arabic dataset for the MBTI personality from Twitter. They gathered tweets from users who had already shared their personality types in their profiles or posts. They obtained data from 796 authors, each with 50 tweets. Salem *et al.* [43] created a website where people can take a Big Five personality test and provide permission to access their Twitter accounts. They collected a sample of 92 Egyptian dialect users. They used a questionnaire comprising 25 randomly chosen questions from a well-defined personality inventory. This dataset was employed in a recent study [44].

As previously stated, the large number of Arabic speakers implies the availability of a massive amount of data that can be processed and exploited as well as a significant population to study in terms of personality. However, the number of studies conducted to date is insufficient. Given this context, we concentrated our efforts on analyzing personality from Arabic text to advance studies in this field.

The first requirement is to have an Arabic dataset. After thorough research, we found none for standard Arabic. Therefore, the contributions of this work are: i) the creation of a dataset of MSA text annotated with the Big Five personality traits, and ii) the evaluation of the dataset, in which we tried some basic text preprocessing techniques, traditional classifiers, and regressors to predict personality classes and scores. We obtained promising results by using a corpus of 267 entries.

The remainder of this paper is organized as follows. The following section presents the method used to construct our dataset, a description of the obtained corpus, and the prediction method comprising text processing, classification, and regression. In the third section, additional details of the experiments and their outcomes are presented. Finally, we conclude the paper with a discussion and conclusion.

2. METHOD

To create a personality classification system that utilizes MSA text, a dataset is required. However, due to the scarcity of MSA and Arabic datasets in general, a new dataset has been developed. Drawing inspiration from previous research, particularly [3], the process of creating this dataset involved two main steps: data gathering and data cleaning. The following subsections detail each step and describe the data collected. In the prediction process subsection, we explain how the corpus is evaluated using a natural language processing (NLP) workflow for text classification. This workflow respects the general Arabic NLP process [45], as it includes text preprocessing, feature extraction, and prediction using machine learning algorithms. The proposed process is described in Figure 2.

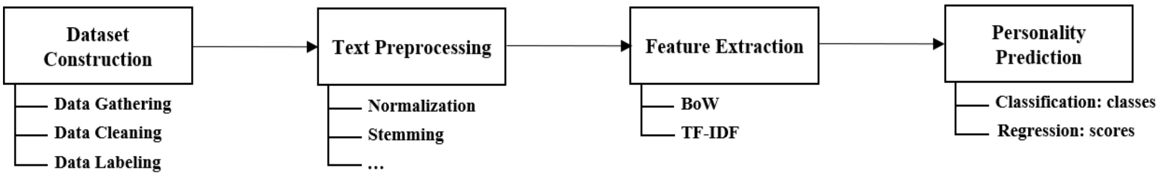

Figure 2. The proposed process to create and evaluate the dataset

2.1. Dataset construction

2.1.1. Data gathering

To gather data, we conceived a form in Arabic and integrated it into a website that we hosted online [46]. This form is divided into three parts. The first concerns demographic information, such as sex, age range, occupation, level of study, and specialty. The second part consists of the Big Five inventory proposed by John, Donahue, and Kentle in 1991 [47]. We used the Arabic version translated by Al Ali and Al Ansari in 2018 [48]. It includes 44 items measured on a 5-point Likert scale (5=I strongly agree, 1=I strongly disagree). The last part of the form is a text area in which people are requested to write freely about anything. We proposed some subjects to help them, such as university experience, work experience, day/week events, and emotions. In this part, people are restricted to writing in Arabic only, and they have to write at least 30 words. The form cannot be submitted if any answer is missing or if a condition is not fulfilled. All the participants were anonymous, which ensured the confidentiality of their responses. We shared the website through the university's email network, Facebook groups, and one-on-one sharing.

2.1.2. Data cleaning

After collecting the data, we manually reviewed them to distinguish any abnormal text entries. We deleted any answers that contained random letters, text copied from the website, text with only repeated words, or dialect text. Subsequently, we removed duplicate entries that were detected automatically.

2.1.3. Dataset labeling

The personality scores were derived from the participants' responses, and each trait falls within a specific range of scores: AGR (9-45), CON (9-45), EXT (8-40), NEU (8-40), and OPN (10-50). The scores for each trait were determined using a set of questions, Q, evaluated using a 5-point Likert scale (R denotes reverse-scored items:1=I strongly agree, 5=I strongly disagree):

- Agreeableness: *Q2_R, Q7, Q12_R, Q17, Q22, Q27_R, Q32, Q37_R, Q42*
- Conscientiousness: *Q3, Q8_R, Q13, Q18_R, Q23_R, Q28, Q33, Q38, Q43_R*
- Extraversion: *Q1, Q6_R, Q11, Q16, Q21_R, Q26, Q31_R, Q36*
- Neuroticism: *Q4, Q9_R, Q14, Q19, Q24_R, Q29, Q34_R, Q39*
- Openness: *Q5, Q10, Q15, Q20, Q25, Q30, Q35_R, Q40, Q41_R, Q44*

We divided each personality dimension into two classes using the median value of the participants' scores. Data entries were labeled as high (Yes/1) for a trait if the score was above the median or low (No/0) if the score was equal to or less than the median. The labeling process resulted in binary classes for each trait.

2.1.4. Dataset description

Over two months, we received 300 answers on our form on the website. After undergoing the cleaning phase, we retained 267 entries in our dataset that we named "MSAPersonality". There were 169 female and 98 male participants of different age ranges. The details are listed in Table 1. Regarding educational background, 253 participants were higher education students/graduates, four had high school degrees, three did not complete their high school, and seven were not specified. Table 2, Figure 3, and Figure 4 show some statistics regarding the personality scores and classes in our dataset. As for the text provided by the participants, Table 3 shows the statistics on the number of characters, words, and sentences.

Table 1. Dataset demographical statistics

Age Group	Female	Male	Total
15-17	1	1	2
18-24	45	17	62
25-34	89	42	131
35-44	24	17	41
45-54	7	14	21
55-64	3	6	9
65+	0	1	1
Total	169	98	267

Table 2. Statistics on personality traits scores in the dataset

	AGR	CON	EXT	NEU	OPN
Min	22	18	11	8	20
Max	45	45	38	40	50
Average	35.31	33.21	25.24	24.09	37.43
Std deviation	4.20	5.00	4.93	6.73	5.34

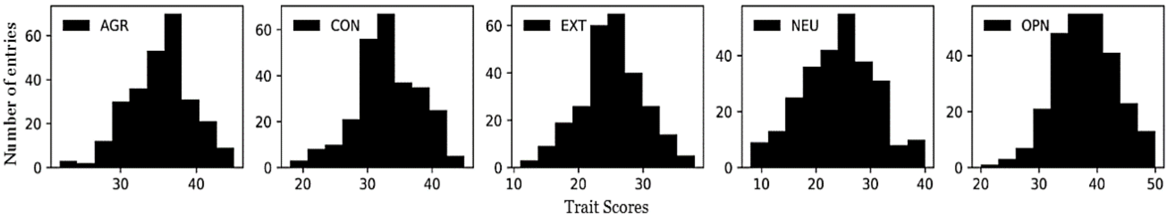

Figure 3. Distribution of personality traits' scores in the dataset

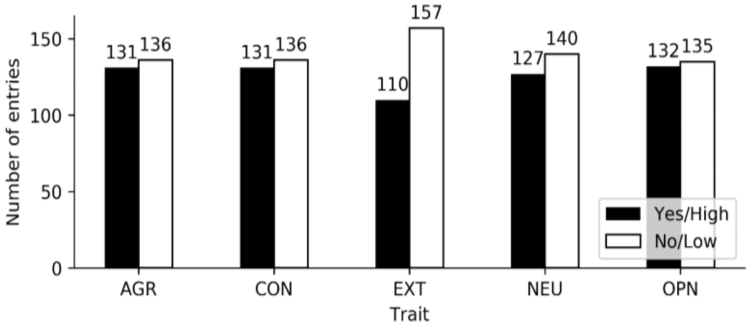

Figure 4. Number of samples for each personality trait class (high/low)

Table 3. Statistics about the text in the dataset

	Characters	Words	Sentences
Count	82151	15121	554
Mean	307.68	56.63	2.07
Minimum	156	30	1
Maximum	1556	283	15

2.2. Prediction process

2.2.1. Text preprocessing

Before building a prediction model, we preprocessed the text using two different methods after a simple cleaning that consisted of removing escape characters ‘\n\r’. The first method (PPM1) involved basic text preprocessing techniques chosen from a study [49] that investigated their impact on Arabic text. We applied the following steps: we removed digits, letters repeated more than twice, diacritics (*tachkeel*), and *Kashida*. In addition, we normalized the text by changing *Tae* (ة) to *Hae* (ه), and *Alif Maksoura* (ى) to *Yae* (ي).

The second method (PPM2) is based on the study [50] in the field of sentiment analysis that was chosen from a review on sentiment analysis in Arabic [51] (the article used MSA and had high accuracy). In addition to removing digits, letters repeated more than twice, diacritics (*tachkeel*), and *Kashida*, this method also removes stop words. Following this, we applied stemming, which removes prefixes and suffixes from words. The examples are listed in Table 4.

Table 1. Example of text preprocessing techniques applied to the original text

Step	Text
Original	حياتي في حالة فوضى ، لا أعرف الطريق الذي يجب أن أسلكه ، أريد أن أكون شخصًا ناجحًا لكنني لم أجد بعد المكان الذي أشعر فيه بالراحة ، أنا متشائم وأريد التغييير للافضل.
Translation	My life is a mess, I don't know which way to go, I want to be a successful person but I haven't yet found the place where I feel comfortable, I'm pessimistic and I want to change for the better.
Removing digits, letters repeated more than twice, diacritics, and Kashida	حياتي في حالة فوضى ، لا أعرف الطريق الذي يجب أن أسلكه ، أريد أن أكون شخصا ناجحا لكنني لم أجد بعد المكان الذي أشعر فيه بالراحة ، أنا متشائم وأريد التغيير للافضل.
Normalization	حياتي في حاله فوضي ، لا أعرف الطريق الذي يجب أن أسلكه ، أريد أن أكون شخصًا ناجحًا لكنني لم أجد بعد المكان الذي أشعر فيه بالراحه ، أنا متشائم وأريد التغييير للافضل.
Remove stop words	حياتي حالة فوضى ، أعرف الطريق يجب أسلكه ، أريد أكون شخصا ناجحا لكنني أجد المكان أشعر بالراحة ، متشائم وأريد التغييير للافضل .
Stemming	حيا في حال فوضى ، لا عرف طريق الذي جب أن سلك ، ريد أن كون شخص ناجح لكن لم جد بعد مكان الذي شعر في راح ، أنا متشائم ريد تغييير أفضل .

2.2.2. Features extraction

We used two vectorization methods (applied separately) to transform the text from natural language to machine language. The first is bag of words (BoW), a technique in which a text document is converted into a vector of word occurrence counts [52]. The second is term frequency - inverse document frequency (TF-IDF). This represents the term frequency-inverse document frequency. It is a twist on BoW that provides a normalized count, where each word count is divided by the number of documents in which the word appears [52]. For both methods, we varied the n-gram parameter between the unigrams, bigrams, tri-grams, and their combination.

2.2.3. Classification

We have five independent Big Five personality traits. For each, there are two classes: yes (high level) or no (low level). Binary classifiers were built for each trait. We tested four traditional classifiers that are widely used in APPR [7]:

- Multinomial naive bayes (MNB): This is a variation of naive Bayes, which is a classification technique based on Bayes' theorem with the assumption of independence among predictors [53].
- MultiLayer perceptron (MLP): It is a model composed of a system of simple interconnected neurons and represents a non-linear mapping between an input vector and an output vector. It can model highly nonlinear functions and be trained to accurately generalize given new, unseen data [54].
- Support vector machine (SVM): It creates hyperplanes in n-dimensional space that can separate examples of different classes [53].
- Random forest (RF): It consists of a combination of tree predictors, where each tree is generated using a random vector and each votes for the most popular class [53].

2.2.4. Regression

For the regression, we used the regressor versions of MLP (MLPR), SVM (SVR), and RF (RFR). For MNB, we replaced it with the K-nearest neighbors regressor (KNNR). k nearest neighbors (KNN) essentially relies on a basic assumption that observations with similar characteristics will tend to have similar outcomes. KNN assigns a predicted value to a new observation based on plurality or the mean of its "k nearest neighbors" in the training set [53].

3. RESULTS AND DISCUSSION

3.1. Experimental setup

The experiments were carried out on a virtual machine running Ubuntu 21.04 LTS and equipped with 8 Intel® Xeon(R) CPU E5-2637 v4 @ 3.50 GHz processors and 64 GB RAM. The implementations were performed using Python in the Jupyter lab [55]. We used CAMeL Tools for most of the Arabic language processing operations [56] and scikit-learn for machine learning functions [57]. During model development, we first split the data into training (70%) and test (30%) sets, and then performed a 5-fold cross-validation on the training data. To optimize the performance of each classifier/regressor, we tuned different hyperparameters using the Python grid search method. Accuracy and root mean square error (RMSE) are the evaluation metrics used for classification and regression, respectively.

3.2. Experimental results

In the experimentation, we tried different preprocessing methods (including no preprocessing), feature extraction techniques, and prediction algorithms. We procured around 240 experiences (scenarios) for each trait. Here, we report the best results only in terms of preprocessing methods and machine learning algorithms with suitable evaluation metrics (accuracy for classification and RMSE for regression).

Our results were compared to a baseline calculated using the zero-rule method. For classification, the rule is to predict the most common class value in the training dataset. As for regression, it predicts the mean of the output values observed in the training data. For simplicity, the following abbreviations will be used in the following tables: No PreProcess (NoPP) and PPM1 and PPM2 for the two methods of preprocessing reported previously.

3.2.1. Classification results

Table 5 lists the results of the classification experiments. We can see that, generally, our classifiers have accuracies above the baseline for most traits. Exceptionally, EXT had the majority of the results equal to the baseline. The only case under the baseline was the OPN with the MNB classifier and NoPP. The best results were obtained with the MNB classifier for CON, EXT, NEU, and OPN traits and with MLP for AGR. Furthermore, in these cases, either the PPM1 or PPM2 preprocessing method outperformed NoPP.

Table 5. Classification Accuracy results with different classifiers and text preprocessing (values in bold are the best results and values in italics are the results less than the baseline)

Classifier		Baseline	MLP	MNB	RF	SVM
Trait	PreProcess					
AGR	NoPP	50.62	54.32	54.32	51.85	54.32
	PPM1		58.02	55.56	51.85	55.56
	PPM2		53.09	56.79	51.85	56.79
CON	NoPP	50.62	62.96	66.67	54.32	61.73
	PPM1		65.43	67.9	50.62	61.73
	PPM2		51.85	55.56	51.85	54.32
EXT	NoPP	59.26	59.26	59.26	59.26	59.26
	PPM1		59.26	59.26	59.26	59.26
	PPM2		60.49	62.96	59.26	59.26
NEU	NoPP	51.85	59.26	62.96	51.85	59.26
	PPM1		61.73	64.2	51.85	59.26
	PPM2		58.02	56.79	51.85	54.32
OPN	NoPP	50.62	51.85	*49.38*	50.62	53.09
	PPM1		50.62	50.62	51.85	53.09
	PPM2		62.96	64.2	58.02	50.62

3.2.2. Regression results

Table 6 presents the results of the regression experiments. For the AGR, EXT, NEU, and OPN traits, the KNN regressor produced the best RMSE results, whereas the MLP regressor produced the best results for CON. Most of the results are better than the baseline even so slightly. However, EXT and NEU had results that exceeded the baseline.

Table 6. Regression RMSE results with different regressors and text preprocessing (values in bold are the best results and values in italics are the results less than the baseline)

Regressor		Baseline	KNNR	MLPR	RFR	SVR
Trait	PreProcess					
AGR	NoPP	3.9500	3.8351	3.9004	3.9057	3.9473
	PPM1		3.8650	3.9134	3.8697	3.9473
	PPM2		3.9024	3.8971	3.8801	3.9475
CON	NoPP	5.3854	5.3696	5.3728	5.3658	5.3729
	PPM1		5.3420	5.3727	5.3716	5.3724
	PPM2		5.3483	5.2244	5.3108	5.3766
EXT	NoPP	4.7660	4.7209	*4.7687*	4.7396	4.7562
	PPM1		4.7066	*4.7751*	4.7558	4.7649
	PPM2		4.6781	*4.8000*	4.7351	4.7602
NEU	NoPP	7.2000	*7.2102*	*7.2054*	7.1844	*7.2011*
	PPM1		7.1927	*7.2059*	7.1903	*7.2004*
	PPM2		7.1768	7.1875	7.1877	7.1902
OPN	NoPP	4.9292	4.9008	4.9118	4.8972	4.8989
	PPM1		4.9012	4.9045	4.9074	4.8988
	PPM2		4.8622	4.8938	4.8770	4.8734

3.3. Discussion

The primary objective of this research was to develop a labeled dataset for MSA texts to study personality recognition in Arabic. The two other studies on Arabic focused on the dialect; the first study [34] collected MBTI personality traits from social media (where users primarily utilize Arabic dialects to express themselves [58]), and the second study [43] gathered Big Five traits and dialect text for Egyptian Twitter users. In terms of the dataset size, 267 entries were available. This is acceptable and comparable to non-English datasets created using the same method, such as a study of modern Greek [40] with 382 entries, and a study of Brazilian Portuguese [29] with 110 participants.

Prediction experiments were conducted to evaluate the corpus and its potential. However, we did not place a high priority on optimizing the performance obtained. The classification results were deemed to be satisfactory. Overall, they were better than the baseline values, with accuracies ranging from 58% for AGR to 67.9% for CON. These results are comparable to some of the state-of-the-art studies reported in [59]. However, we obtained less satisfactory results for the regression problem as they improved only slightly from the baseline. This is similar to the results for other datasets with relatively similar sizes [40].

These results are understandable, and the small text lengths (ranging from 30 to 283 words, with an average of 56.63 words) and basic feature set used may have influenced them. It is worth noting that the implementation of various text preprocessing techniques contributed to the improvement of the results. These findings demonstrate the potential of the proposed corpus. Further research is needed, including the exploration of more suitable preprocessing techniques, experimentation with alternative feature extraction or machine learning algorithms, and most importantly, the expansion of the dataset size.

A comprehensive corpus of MSA holds great significance for the advancement of personality recognition. It not only facilitates a better understanding of the Arabic-speaking population but also enables the customization of services to cater to their preferences. This customization leads to improved interactions and the acquisition of more valuable feedback from the community.

4. CONCLUSION

This paper presented a newly created dataset of Modern Standard Arabic text annotated with the Big Five traits for Automatic Personality Recognition. To the best of our knowledge, this is one of the first studies to address this issue, with previous Arabic research focusing on the MBTI personality or a specific Arabic dialect. The data were gathered online through a website. It has a form with a Big Five inventory and space to write in Arabic. After filtering, we obtained 267 valid entries. We conducted basic prediction experiments to evaluate the dataset and its effectiveness in the classification and regression of personality traits. The results are above chance level, and they are promising, especially for classification. Despite its limitations, our dataset has the potential for use in APR. In the future, more sophisticated and appropriate techniques for text preprocessing, feature extraction, and machine learning will be investigated, and the dataset size will increase.

REFERENCES

[1] D. C. Funder, “Personality,” *Annual Review of Psychology*, vol. 52, no. 1, pp. 197–221, 2001, doi: 10.1146/annurev.psych.52.1.197.

[2] P. Costa and R. McCrae, “The NEO personality inventory manual,” *Psychological Assessment Resources, Odessa*, 1985.

[3] 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.

[4] J. B. Hirsh and J. B. Peterson, “Personality and language use in self-narratives,” *Journal of Research in Personality*, vol. 43, no. 3, pp. 524–527, Jun. 2009, doi: 10.1016/j.jrp.2009.01.006.

[5] 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.

[6] A. Vinciarelli and G. Mohammadi, “A survey of personality computing,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, Jul. 2014, doi: 10.1109/TAFFC.2014.2330816.

[7] K. Chraibi, I. Chaker, and A. Zahi, “Automatic personality prediction: a systematic mapping study,” in *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*, Dec. 2020, pp. 2053–2060, doi: 10.1109/SSCI47803.2020.9308479.

[8] J. Lane, “The 10 most spoken languages in the world,” *Babbel Magazine*, https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world (accessed Apr. 01, 2022).

[9] D. J. Stillwell and M. Kosinski, “myPersonality project: Example of successful utilization of online social networks for large-scale social research,” *American Psychology*, vol. 59, no. 2, pp. 93–104, 2004.

[10] E. Stammatatos *et al.*, “Overview of the 3rd author profiling task at PAN 2015,” in *CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*, 2015, vol. 1391, no. 31, pp. 898–927.

[11] D. Xue *et al.*, “Deep learning-based personality recognition from text posts of online social networks,” *Applied Intelligence*, vol. 48, no. 11, pp. 4232–4246, Jun. 2018, doi: 10.1007/s10489-018-1212-4.

[12] 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, Sep. 2018, doi: 10.1007/s10796-017-9793-8.

[13] E. Castillo, O. Cervantes, and D. Vilariño, “Author profiling using a graph enrichment approach,” *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 5, pp. 3003–3014, May 2018, doi: 10.3233/JIFS-169485.

[14] M. Giménez, R. Paredes, and P. Rosso, “Personality recognition using convolutional neural networks,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10762, Springer International Publishing, 2018, pp. 313–323.

[15] Z. Wen, J. Cao, Y. Yang, H. Wang, R. Yang, and S. Liu, “DesPrompt: personality-descriptive prompt tuning for few-shot personality recognition,” *Information Processing and Management*, vol. 60, no. 5, Sep. 2023, doi: 10.1016/j.ipm.2023.103422.

[16] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, “Bottom-up and top-down: predicting personality with psycholinguistic and language model features,” in *2020 IEEE International Conference on Data Mining (ICDM)*, Nov. 2020, pp. 1184–1189, doi: 10.1109/ICDM50108.2020.00146.

[17] K. El-Demerdash, R. A. El-Khoribi, M. A. Ismail Shoman, and S. Abdou, “Deep learning based fusion strategies for personality prediction,” *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 47–53, Mar. 2022, doi: 10.1016/j.eij.2021.05.004.

[18] 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.

[19] W. R. Wright and D. N. Chin, “Personality profiling from text: Introducing part-of-speech N-grams,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8538, Springer International Publishing, 2014, pp. 243–253.

[20] C. Solinger, L. Hirshfield, S. Hirshfield, R. Friendman, and C. Leper, “Beyond Facebook personality prediction: a multidisciplinary approach to predicting social media users’ personality,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8531, Springer International Publishing, 2014, pp. 486–493.

[21] V. Lytvyn, P. Pukach, I. Bobyk, and V. Vysotska, “The method of formation of the status of personality understanding based on the content analysis,” *Eastern-European Journal of Enterprise Technologies*, vol. 5, no. 2–83, pp. 4–12, Oct. 2016, doi: 10.15587/1729-4061.2016.77174.

[22] Y. Hernández, C. A. Peña, and A. Martínez, “Model for personality detection based on text analysis,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11289, Springer International Publishing, 2018, pp. 207–217.

[23] M. Komisin and C. Guinn, “Identifying personality types using document classification methods,” in *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference*, 2012, pp. 232–237.

[24] 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*, Oct. 2015, pp. 9–14, doi: 10.1109/WOCC.2015.7346106.

[25] R. Gao, B. Hao, S. Bai, L. Li, A. Li, and T. Zhu, “Improving user profile with personality traits predicted from social media content,” in *RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems*, Oct. 2013, pp. 355–358, doi: 10.1145/2507157.2507219.

[26] X. Liu and T. Zhu, “Deep learning for constructing microblog behavior representation to identify social media user’s personality,” *PeerJ Computer Science*, vol. 2016, no. 9, Sep. 2016, doi: 10.7717/peerj-cs.81.

[27] P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha, “25 tweets to know you: A new model to predict personality with social media,” *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, vol. 11, no. 1, pp. 472–475, May 2017, doi: 10.1609/icwsm.v11i1.14963.

[28] C. Li, J. Wan, and B. Wang, “Personality prediction of social network users,” in *Proceedings - 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, DCABES 2017*, Oct. 2017, vol. 2018-Septe, pp. 84–87, doi: 10.1109/DCABES.2017.25.

[29] V. G. Dos Santos, I. Paraboni, and B. B. C. Silva, “Big five personality recognition from multiple text genres,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10415, Springer International Publishing, 2017, pp. 29–37.

[30] 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, pp. 6877–6888, 2018.

[31] J. Shen, O. Brdiczka, and J. Liu, “Understanding email writers: Personality prediction from email messages,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7899, Springer Berlin Heidelberg, 2013, pp. 318–330.

[32] 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, Sep. 2016, doi: 10.1007/s13278-016-0383-4.

[33] T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, “Personality prediction system from Facebook users,” *Procedia Computer Science*, vol. 116, pp. 604–611, 2017, doi: 10.1016/j.procs.2017.10.016.

[34] N. Alsadhan and D. Skillicorn, “Estimating personality from social media posts,” in *IEEE International Conference on Data Mining Workshops, ICDMW*, Nov. 2017, vol. 2017-Novem, pp. 350–356, doi: 10.1109/ICDMW.2017.51.

[35] A. V Kunte and S. Panicker, “Using textual data for personality prediction: a machine learning approach,” in *2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019*, Nov. 2019, pp. 529–533, doi: 10.1109/ISCON47742.2019.9036220.

[36] W. Maharani and V. Effendy, “Big Five personality prediction based in Indonesian tweets using machine learning methods,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1973–1981, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1973-1981.

[37] M. A. Kosan, H. Karacan, and B. A. Urgen, “Personality traits prediction model from Turkish contents with semantic structures,” *Neural Computing and Applications*, vol. 35, no. 23, pp. 17147–17165, Apr. 2023, doi: 10.1007/s00521-023-08603-z.

[38] A. M. Dos Santos *et al.*, “Words similarities on personalities: a language-based generalization approach for personality factors recognition,” *IEEE Access*, vol. 11, pp. 29823–29836, 2023, doi: 10.1109/ACCESS.2023.3261339.

[39] S. Nowson and J. Oberlander, “Identifying more bloggers: Towards large scale personality classification of personal weblogs,” *International Conference on Weblogs and Social Media*, ICWSM-2007, 2007.

[40] V. Komianos, E. Moustaka, M. Andreou, E. Banou, S. Fanarioti, and K. L. Kermanidis, “Predicting personality traits from spontaneous modern Greek text: Overcoming the barriers,” in *IFIP Advances in Information and Communication Technology*, vol. 382, Springer Berlin Heidelberg, 2012, pp. 530–539.

[41] A. C. E. S. Lima and L. N. de Castro, “A multi-label, semi-supervised classification approach applied to personality prediction in social media,” *Neural Networks*, vol. 58, pp. 122–130, Oct. 2014, doi: 10.1016/j.neunet.2014.05.020.

[42] L. Gou, J. Mahmud, E. M. Haber, and M. X. Zhou, “PersonalityViz: a visualization tool to analyze people’s personality with social media,” in *International Conference on Intelligent User Interfaces, Proceedings IUI*, Mar. 2013, pp. 45–46, doi: 10.1145/2451176.2451191.

[43] M. S. Salem, S. S. Ismail, and M. Aref, “Personality traits for Egyptian twitter users dataset,” in *ACM International Conference Proceeding Series*, Apr. 2019, pp. 206–211, doi: 10.1145/3328833.3328851.

[44] H. Elzayady, M. S. Mohamed, K. M. Badran, and G. I. Salama, “A hybrid approach based on personality traits for hate speech detection in Arabic social media,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 1979–1988, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1979-1988.

[45] G. Bourahouat, M. Abourezq, and N. Daoudi, “Systematic review of the Arabic natural language processing: challenges, techniques and new trends,” *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 3, pp. 1333–1343, 2023.

[46] “Personality test,” (in Arabic) LSIA, http://personality.rf.gd/ (accessed Jun. 01, 2022).

[47] O. P. John, E. M. Donahue, and R. L. Kentle, “Big five inventory,” *Journal of Personality and Social Psychology*, 1991.

[48] T. B. Al Ali 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. 2, pp. 167–203, Jun. 2018, doi: 10.12785/jeps/190206.

[49] Y. Albalawi, J. Buckley, and N. S. Nikolov, “Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media,” *Journal of Big Data*, vol. 8, no. 1, Jul. 2021, doi: 10.1186/s40537-021-00488-w.

[50] M. A. Sghaier and M. Zrigui, “Sentiment analysis for Arabic e-commerce websites,” *2016 International Conference on Engineering & MIS (ICEMIS)*, Agadir, Morocco, 2016, pp. 1-7, doi: 10.1109/ICEMIS.2016.7745323.

[51] M. E. M. Abo, R. G. Raj, and A. Qazi, “A review on Arabic sentiment analysis: state-of-the-art, taxonomy and open research challenges,” *IEEE Access*, vol. 7, pp. 162008–162024, 2019, doi: 10.1109/ACCESS.2019.2951530.

[52] A. Zheng and A. Casari, “Feature engineering for machine learning: principles and techniques for data scientists,” *O'Reilly*, 2018.

[53] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: a survey,” *Information*, vol. 10, no. 4, Apr. 2019, doi: 10.3390/info10040150.

[54] H. Taud and J. F. Mas, “Multilayer perceptron (MLP),” in *Lecture Notes in Geoinformation and Cartography*, Springer International Publishing, 2018, pp. 451–455.

[55] 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.

[56] O. Obeid *et al.*, “CAMeL tools: an open source python toolkit for Arabic natural language processing,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, May 2020, pp. 7022–7032.

[57] Fabian P. *et al.*, “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[58] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, “A review of sentiment analysis research in Arabic language,” *Future Generation Computer Systems*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/j.future.2020.05.034.

[59] S. Mushtaq and N. Kumar, “Text-based automatic personality recognition: recent developments,” in *Lecture Notes in Networks and Systems*, vol. 421, Springer Nature Singapore, 2023, pp. 537–549.

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.

Younes Dhassi received the higher education degrees and Ph.D. degrees in computer sciences from USMBA-Fez University in 2018, respectively. He is currently an assistant professor with the Department of Computer Science, Faculty of Sciences and Technology, USMBA-Fez University. His research interests are in the areas of computer vision, particularly in pattern recognition, image and video processing, and data analysis. He can be contacted at email: younes.dhassi@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.