International Journal of Intelligent Computing and Information Science

PREPROCESSING THE EGYPTIAN ARABIC DIALECT FOR PERSONALITY TRAITS PREDICTION

M. S. Salim S.S. Ismail M. Aref
Department of Computer Sciences, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

Marwa_salim@cis.asu.edu.eg sallySaad@gmail.com Mostafa.aref@fcis.asu.edu.eg

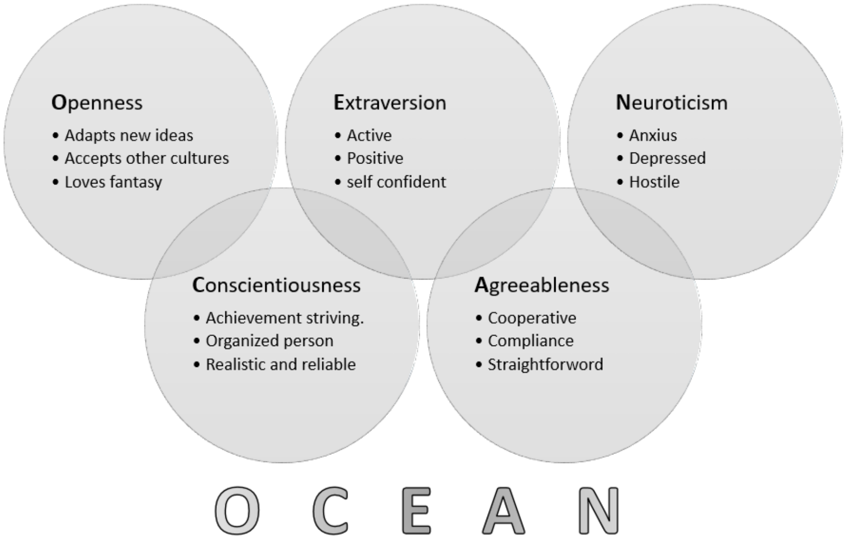
Abstract*: Each individual has his own distinct character, making his own decisions which is based on his personality. Researchers in computer science field have tried to reach a model for extracting personality traits relying on user's profiles on social network sites as an input. Content created by users such as text posts, photos and shared activities in social network sites are considered as a huge source of data. Regarding user-created text, it has been proved that text pre-processing has a great impact if was applied to text before using it in research. In this paper, the effect of pre-processing (stemming and stop word removal) and adding numerical features is tested on the performance of Arabic personality prediction using AraPersonality dataset, which yielded 3.0% and 6.7% overall improvement to baseline experiments in binary representation and multiclass representation respectively.*

Keywords: *Personality Recognition, Social Media, AraPersonality Dataset, Stop Word, Stemming.*

1 Introduction

Psychologists have discovered that there are traits, which differentiate the behaviour of each person from others, but they have differed on how to define them. The most widespread and used way is the five factor model (FFM) or the big five (BF) [4] which is shown in Figure 1. The big five defines personality through five traits named Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (OCEAN). Taking into consideration that more than one trait can be found in the same person. Person's behaviour should be studied in order to estimate his personality traits. Such data can be found in social media feeds of its users. Social network sites are gradually spreading worldwide and its users are significantly increasing. The extraction of personality traits for them is useful in many businesses models such as directed advertising and healthcare such as mood detection.

Figure 1 The Big Five Personality Traits



Arabic is one of the six official languages of the United Nations Organization, which is the native language for 467 million of the world's population. However, less attention is paid to it in the field of scientific research, as it is classified within the most difficult languages in dealing with in scientific researches. Its linguistic material is abundant. A word may have different shapes in writing based on its location in the sentence but must treat these different forms as the same word and a vast collection of words are considered to be stop words so must be ignored. So pre-processing to Arabic text must be carried out before using it in research. This paper discuss the effect of pre-processing on the accuracy of Arabic personality prediction using AraPersonality dataset.

The rest of this paper is organized as follows: Related work is in Section 2. Section 3 explains AraPersonality dataset and the methodology used in this work. The experimental setup and results are presented in section 4. Section 5 presents conclusions and avenues for future research.

2 Related Work

The researchers used social media networks to gather many datasets to predict personality because social media networks are considered to be one of the richest information recourses about its users as shown in Table 1.They used different approaches of the users profiles data to predict personality like what user's written content, profile pictures and other profile information.

Table 1 Personality Prediction Datasets

Name	**Description**	**Number Of Users**	**Used**	**Language**
AraPersonality [16]	Dataset contains users profile data, twitter feeds and personality scores from twitter	92	[16]	Arabic (Egyptian dialect)
MyPersonality [11]	Dataset contains users profile data and personality scores from Facebook	more than 11 million	[1, 5, 7, 9, 19, 20, 22, 23, 27]	English
Whitty et al. [25]	Dataset contains users profile images and personality scores from Facebook and twitter.	207	[25]	None
Bhatti et al. [3]	Dataset contains profile pictures, posts, online engagement, and personality scores from twitter	54,784	[3]	English

JamesPennebaker and Laura King's essay [14]	Dataset contains essay and personality scores from essay	2,467.	[10]	English
Xue et al. [26]	Dataset contains users profiles, micro-blogs and personality scores from sina weibo	994	[26]	Chinese
Wan et al. [24]	Dataset contains users profiles and personality scores from sina weibo	131	[24]	Chinese
Nie et al. [12]	Dataset contains Users information and personality scores from sina Microblog	1792	[12]	Chinese
YouTube personality [6]	Dataset contains 404 users/videos, transcriptions and personality scores from YouTube	404	[2, 7, 8, 17]	English
Twitter [15]	Dataset contains age, tweets, gender and personality scores from twitter	102	[7]	English, Dutch, Spanish and Italian
Tighe et al. [21]	Dataset contains age, gender, tweets and personality Scores from twitter	250	[21]	Filipino and English

Some of researchers focused on what users write. As a preview, Majumder et al. [10] extracted word2vec embedding and Mairesse features as input to different approaches. These approaches are the convolutional neural network (Deep CNN), multiple layer perceptron (MLP) and a polynomial Support Vector Machine (SVM). Xue et al. [27] extracted statistical linguistic, deep semantic features and dictionary-based features feed them into traditional regression algorithms. Carducci et al. [5]used word vector as features to SVM and used dataset from multisource. Varshney et al. [22] and Farnadi et al. [7] used multisource datasets as try to generalize personality recognition through all social media. Alam et al. [1] extracted TF-IDF of unigrams as input to Sequential Minimal Optimization for Support Vector Machine (SMO) with linear kernel, Bayesian Logistic Regression (BLR) and MNB. Verhoeven et al. [23] tried using Ensemble Method for Personality Recognition by using the output of five classifiers (one for each personality trait) based on 250 user from myPersonality and the output of five classifiers (one for each personality trait) trained on the complete essay data. They extracted main feature 2000 most frequent character trigrams as input to SMO. Marwa et al. [16] used tf-idf for n-grams as features to four machine learning algorithm. These algorithms are SVM, Multinomial Naïve Bayes (MNB), KNN and decision tree. For more details, a fully detailed review can be found in [16].

Others focused on personality prediction from profile pictures. Whitty et al. [25] used users, which had both an active Facebook and Twitter account with Binary logistic regression to predict personality. Bhatti et al. [3] extracted colour, image composition, image type, and diversification and Facial Presentation as the input of linear regression and Pearson correlations to predict personality of the USA people using Twitter.

Others focused on personality recognition from different type profile information. Xue et al. [26] extracted profile-based static (gender, address, nickname, etc.), profile-based dynamic (number of followers and followings) and content-based micro-blogs (linguistic features, psychological features) as input to paradigm eight label distribution learning (LDL) algorithms. Wan et al. [24] extracted user behaviour (count, followers count, followings count, time since registration), interaction behaviour (expressions count, topics count and @ mentions) and text features Linguistic Inquiry and Word Count (LIWC) as input to two classification model Logistic regression and Naïve Bayes algorithms. Nie et al. [12] tried semi-supervised learning by the using of unlabelled samples in personality prediction. They extracted summarized features from raw data directly and statistical features as input to local linear semi-supervised regression algorithm. Farnadi et al. [9] extracted LIWC features, Social Network features, time-related features, other features as input to SMO with linear kernel, BLR and MNB sparse modelling.

Farnadi et al. [8] tried combined prediction for all five personality trait scores, instead of dealing with each trait separately by using Multivariate regression techniques. They used Gender, Audio-Video, LIWC, NRC, MRC, SentiStrength and Structured Programming for Linguistic Cue Extraction (SPLICE) as input to five Multivariate regression techniques and Multi-objective random forest (MORF). Alam et al. [2] tried used predicted traits as features by designing a cascaded classification system. They used audio-visual, lexical, part of speech (POS), psycholinguistic (LIWC) and emotional features after make Relief algorithm (feature selection) as input to SMO as classifier with different kernels and they build three models called Maj-5 model, Maj-4 Model, Maj-5-traits model. Sarkar et al. [17] used Audio-visual features (AV), Word statistics features (W), Sentiment features (S), Gender, Text as features. Then they ranked set of features according to traits then selected top ten features as input to logistic regression model with a ridge estimator. Tandera et al. [20] used LIWC2015, SPLICE and SNA as features and they used deep learning algorithm using four architectures, namely LSTM (Long Short Term Memory), MLP, GRU (Gated Recurrent Unit), and CNN 1D (1 -Dimensional Convolutional Neural Network). TADESSE et al. [19] used SNA, SPLICE and LIWC then make Pearson correlation as feature selection then used result as features to SVM, Logistic Regression, Gradient Boosting and XGBoost as machine learning algorithm.

3 Proposed Methodology

By modifying on the methodology used in the baseline research of Arapersonality dataset as shown in Figure 2. Two parts were modified. The first is adding the data pre-processing stage. The second is modifying features extracting. The updates will explain in this section.

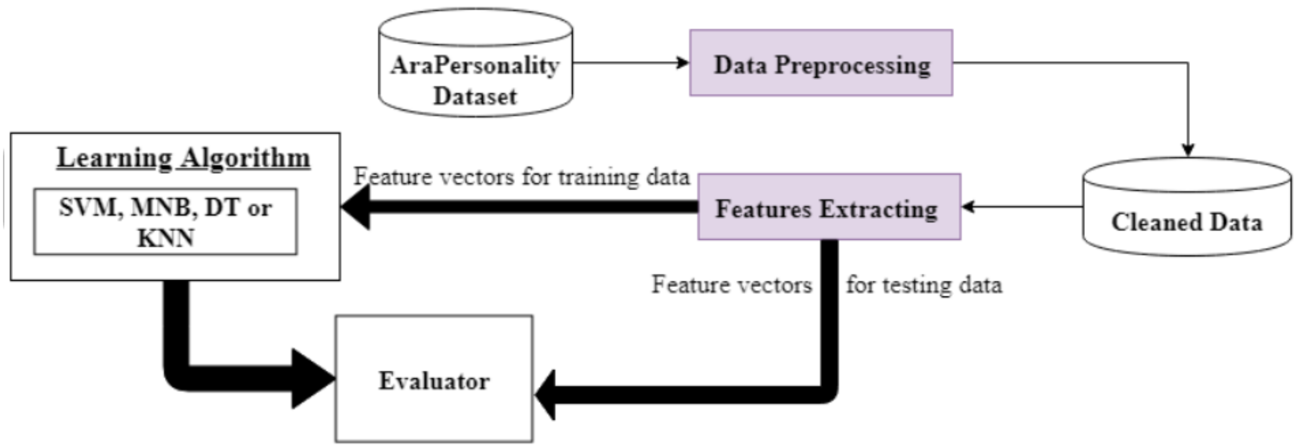


Figure 2 The Used Methodology

3.1 AraPersonality Dataset

This dataset contains the profile data, twitter feed and personality scores based on the big five personality traits of 92 Egyptian twitter users with approximately 3200 tweets per user. The scores of this dataset has two different representation. The first is binary representation such that user's score has value equal zero or one for each trait. The second is multiclass representation such that user's score has value range from one to five.

3.2 Data Pre-processing

Two steps were added in the part of data pre-processing. These steps are stemming and stop words removal as shown in Figure 3.

Figure 3 Data Pre-processing

3.1.1 Stemming

In the Arabic language, converting a word to a verb, adjective, adverb or plural is done by adding some letters to the original word. These additions may be in prefix, suffix or infix. Stemming is to return the word to its original form by removing these additions and preserve its primitive structure. Despite of in some cases the result of the stem may be different from the original word's root, related words usually are mapped to the same stem.

3.1.2 Stop words removal

Stop words are common words used to link sentences to each other and don't add to the meaning of the sentences, these words should be removed as they do not contain important information such as "في" (in), "على" (on), "انت" (you), "من" (of), and similar words. The general characteristics of the stop words are being frequently repeated, general word and its existence alone does not carry meaning.

3.3 Features Extracting

New step was added in this phase, which is numerical features selection as shown in Figure 4.

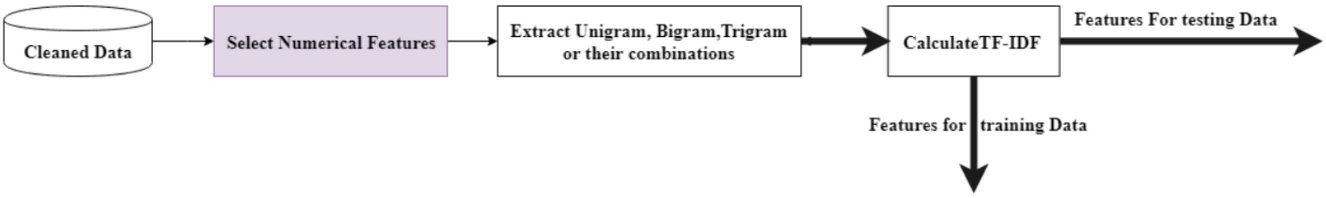


Figure 4 Feature Extracting

3.3.1 Numerical Feature selection

Some numerical features found to have an impact in predicting personality traits that must be identified and others with no impact should be ignored. Therefore, the correlation coefficient between each numerical feature and traits will be used to determine how closely they relate to each other. Features with absolute correlation coefficient greater than 0.05 will be chosen.

The spearman's rank correlation coefficient was used to assess the strength of the relationship between the different numerical features and each of the five personality traits. Spearman's correlation is used as it assesses monotonic relationships (whether linear or not). The correlation coefficient (r) ranges from -1.0 to +1.0. If r is positive means a direct relationship between the variables. If r is negative means an inverse relationship between the variables. If r equal zero means no relationship between variables.

4 Experiment

4.1 Experimental Setting

Stemmer and stop word list are used from [18]. All the experiments were implemented in Python using Scikit Learn [13] using 10 fold cross validation. ASUS laptop computer of model: K556U is used. The processor is Intel® Core™ i7-7500 CPU @ 2.70 GHz (2 cores) and 12GB of RAM.

4.2 Experimental Results

As a result of applying spearman's correlation to select which numerical features will be used, Table 2 and Table 3 show the results for binary and multiclass representations respectively. In these tables, the selected features are written in bold where the absolute of correlation coefficient value is greater than 0.05 for example the results in binary representation for openness and extraversion, all numerical features except number of tweets were used and the results in multiclass representation for agreeableness and extraversion, all numerical features were used.

Table 2 Correlation Result in Binary Representation

Feature Name / Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Number of Tweets	0.047104	**0.152704**	0.019935	0.023781	**0.136601**
Number of Followers	**0.189821**	**0.104202**	**0.199961**	**0.157904**	-0.00117
Number of Following	**0.061052**	-0.0053	**0.088346**	0.035108	0.047683
Number of Favourites	**0.189257**	**0.070365**	**0.095642**	0.013329	**0.091033**
Number of Tweets Per Day	**0.064134**	0.035618	**0.307646**	**0.137451**	-0.01731
Number of Retweet	**0.064273**	**0.20969**	**0.114603**	-0.03624	**0.162971**
Number of Reply	**0.122736**	**0.203426**	**0.173155**	0.002419	**0.157318**

Table 3 Correlation Result in Multiclass Representation

Feature Name / Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Number of Tweets	-0.02215	**-0.14125**	**0.110303**	**0.131105**	**0.255198**
Number of Followers	0.044295	**-0.11198**	**0.265485**	**0.237912**	-0.00193
Number of Following	**-0.07607**	**-0.114**	**0.31136**	**0.231114**	**-0.06933**
Number of Favourites	**0.051999**	**-0.10795**	**0.233279**	**0.172852**	0.031777
Number of Tweets Per Day	**-0.11912**	**-0.08632**	**0.457804**	**0.172177**	0.023823
Number of Retweet	-0.0077	**-0.07164**	**0.142531**	**0.112666**	**0.288936**
Number of Reply	0.001928	0.004039	**0.227616**	**0.173971**	**0.15613**

Table 4 and Table 5 compare f1-score for baseline experiment versus two other models. The first one is the same model used in the baseline experiment after adding pre-processing and relevant numerical features to it. The second model is the best model after running more than 350 experiment. These experiment applied different combinations of features and pre-processing steps per trait. F1-score is shown in bold if it is greater than the baseline model result. An asterisk was added beside the best value per trait in each model. The best value for each trait shows that the performance of openness and neuroticism traits were not improved in binary representation and the performance of conscientiousness and neuroticism traits were not improved in multiclass representation as shown in Figure 5 and Figure 6. The average of best trait values for each model were calculated to measure the performance improvement. The average shows that the first model did not improve the result, but the second one showed improvement 3.0% and 6.7% over the baseline in binary and multiclass representation respectively.

Table 4 F1-Scores for the Three Models in Binary Representation

Trait	**ML**	**Baseline model**	**First Model**	**Second Model**
Openness	DT	0.53	**0.59***	**0.59***

	KNN	0.75*	0.42	0.5
	MN B	0.42	**0.45**	**0.49**
	SVM	0.37	**0.5**	**0.55**
Conscientiousnes s	DT	0.38*	**0.42***	**0.48**
	KNN	0.34	0.31	**0.53***
	MN B	0.21	0.21	**0.25**
	SVM	0.15	**0.31**	**0.31**
Extraversion	DT	0.5	**0.58**	**0.62***
	KNN	0.53*	**0.62***	**0.62**
	MN B	0.35	0.29	0.3
	SVM	0.3	**0.41**	**0.43**
Agreeableness	DT	0.56	0.5	**0.65**
	KNN	0.62*	**0.66***	**0.71***
	MN B	0.56	0.56	0.56
	SVM	0.4	**0.53**	**0.54**
Neuroticism	DT	0.65	0.63*	0.63*
	KNN	0.71*	0.49	0.59
	MN B	0.47	0.47	**0.49**
	SVM	0.53	0.51	**0.61**
Average	**Best**	0.598	0.584	**0.616**

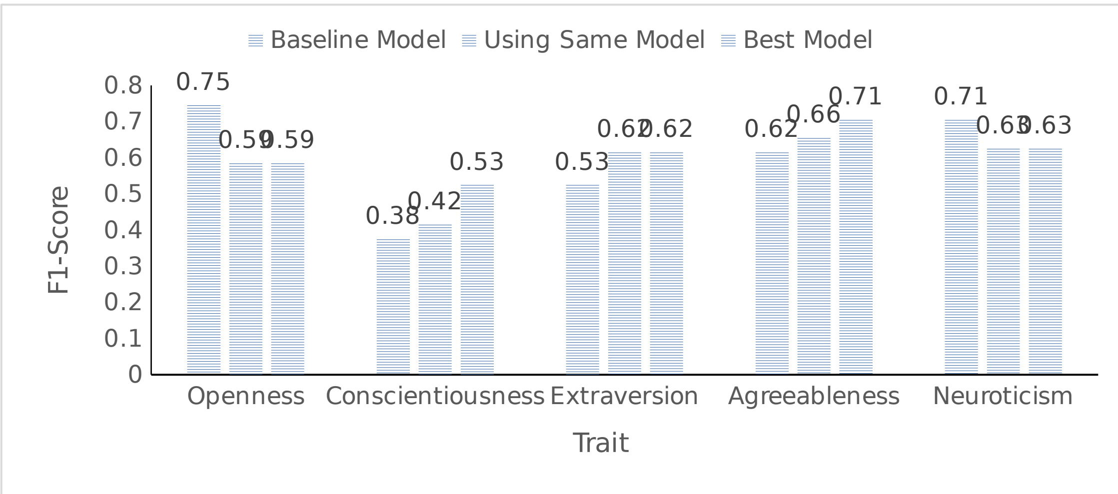


Figure 5

Comparison between The best Results Per Trait in Binary Representation

Table 5 F1-Scores for the Three Models in Multiclass Representation

Trait	ML	Baseline Model	First Model	Second Model
Openness	DT	0.31	0.29*	**0.39***
	KNN	0.32*	0.24	0.26
	MN B	0.28	0.23	0.24
	SVM	0.29	0.21	0.24
Conscientiousnes	DT	0.19	0.15	**0.26***

s	KNN	0.27*	0.19*	0.22
	MNB	0.06	0.06	0.06
	SVM	0.08	0.07	0.07
Extraversion	DT	*0.4*	**0.46***	**0.46***
	KNN	0.4	0.35	**0.43**
	MNB	0.4	0.4	0.4
	SVM	0.41*	0.4	0.4
Agreeableness	DT	0.27	**0.29**	**0.29**
	KNN	0.22	**0.36***	**0.36***
	MNB	0.27	**0.31**	**0.31**
	SVM	0.32*	0.31	0.31
Neuroticism	DT	0.31*	0.25*	0.27*
	KNN	0.22	0.15	**0.25**
	MNB	0.15	0.15	0.15
	SVM	0.14	**0.15**	**0.15**
Average	**Best**	0.326	0.31	**0.348**

Table 6 and Table 7 show which n-gram features were used for binary and multiclass representations in the best model such that text features used in first model are the same as features used in baseline experiments. As example openness trait with decision tree used only unigram as text features in binary and multiclass representations. Table 8 and Table 9 show which pre-processing steps (stemming and stop words removal) were used in the mentioned models in both binary and multiclass representations respectively. These tables show that models used different combination of stemming and stop word removal.

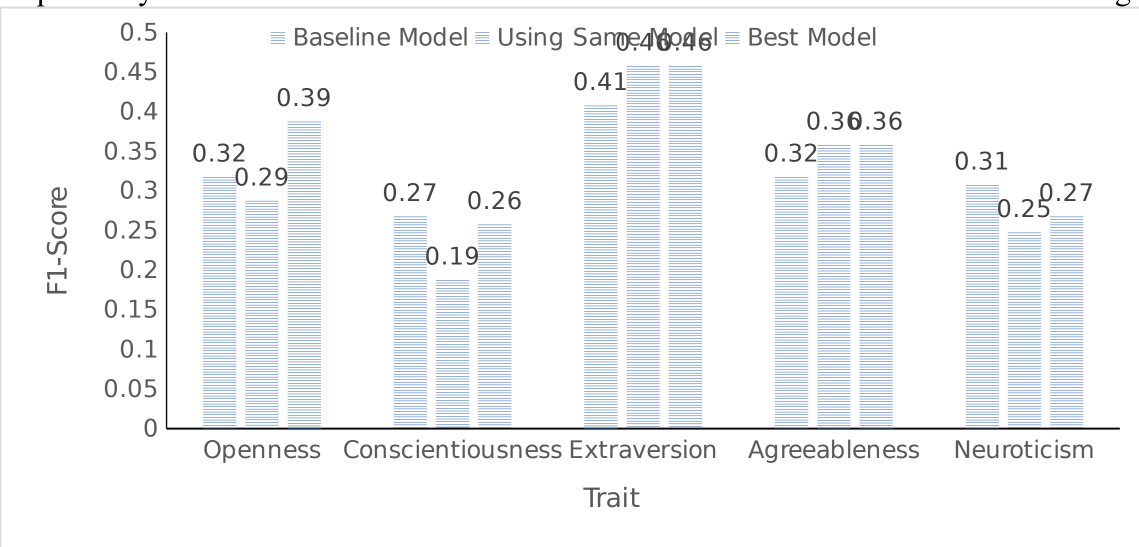


Figure 6 Comparison between The best Results Per Trait in Multiclass Representation

Table 6 N-grams Used in Binary Representation (Second Model)

Trait	ML	1g	2g	3g
Openness	DT	√		
	KNN		√	√
	MNB			√
	SVM	√	√	√
Conscientiousness	DT	√		
	KNN	√	√	√
	MNB			√
	SVM	√		
Extraversion	DT			√
	KNN and SVM			√
	MNB	√		
Agreeableness	DT	√		
	KNN	√		
	MNB			√
	SVM		√	√
Neuroticism	DT	√	√	
	KNN		√	
	MNB			√
	SVM	√	√	

Table 7 N-grams Used in Multiclass Representation (Second Model)

Trait	ML	1g	2g	3g
Openness	DT	√		
	KNN and SVM			√
	MNB		√	√
Conscientiousness	DT	√		
	KNN		√	
	MNB	√		
	SVM	√		
Extraversion	DT			√
	KNN		√	
	MNB	√		
	SVM	√		
Agreeableness	DT	√	√	√
	KNN			√
	MNB	√		
	SVM	√		
Neuroticism	DT	√	√	
	KNN	√	√	
	MNB and SVM	√		

Table 8 Preprocessing Used in Binary Representation

Model	**Trait**	**ML**	**Stem**	**Stop**
First Model	Openness	DT and KNN	✓	
		MNB and SVM		✓
	Conscientiousness	DT and SVM	✓	
		KNN and MNB	✓	✓
	Extraversion	DT and KNN		✓
		MNB	✓	✓
		SVM	✓	
	Agreeableness	DT and MNB		✓
		KNN	✓	✓
		SVM	✓	
	Neuroticism	DT and KNN		✓
		MNB	✓	✓
		SVM	✓	
Second Model	Openness	DT and MNB	✓	✓
		KNN and SVM		✓
	Conscientiousness	DT, KNN and MNB	✓	✓
		SVM	✓	
	Extraversion	DT and MNB	✓	
		KNN and SVM		✓
	Agreeableness	DT	✓	
		KNN	✓	✓
		MNB and SVM		✓

	Neuroticism	DT	√	√
		KNN, MNB and SVM		√

Table 9 Preprocessing Used in Multiclass Representation

Model	**Trait**	**ML**	**Stem**	**Stop**
First Model	Openness	DT, KNN and SVM	√	√
		MNB	√	
	Conscientiousness	DT, KNN and MNB	√	√
		SVM	√	
	Extraversion	All algorithm	√	√
	Agreeableness	DT and MNB		√
		KNN and SVM	√	
	Neuroticism	DT		√
		KNN, MNB and SVM	√	√
Second Model	Openness	DT		√
		KNN and SVM	√	
		MNB	√	√
	Conscientiousness	DT		√
		KNN and SVM	√	√
		MNB	√	
	Extraversion	DT, KNN and MNB	√	√
		SVM	√	
	Agreeableness	DT and MNB		√
		KNN and SVM	√	
	Neuroticism	DT	√	

		KNN, MNB and SVM	ٮا	ٮا

5 Conclusion and future work

A model for extracting personality traits relying on user's profiles on social network sites as an input has been studied. Content created by users such as text posts, photos and shared activities in social network sites are considered as a huge source of data. In this paper, the effect of pre-processing for Egyptian dialect users and adding numerical features to predict personality traits is presented. The pre-processing consists of stemmer and stop words removal. The best model in binary representation showed an improvement of 3.0% over the baseline model, while the best model in the multiclass representation showed an improvement of 6.7% over the baseline model. Next steps is planned to try different types of features. Extend feature selection techniques that can improve the result.

References

[1] Alam, Firoj and Stepanov, Evgeny A and Riccardi, G. Personality Traits Recognition on Social Network - Facebook. *2013 International Conference on Weblogs and Social Media, ICWSM 2013 Workshop*. 6–9.

[2] Alam, F. and Riccardi, G. 2014. Predicting Personality Traits using Multimodal Information. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition - WCPR '14*. (2014), 15–18. DOI:https://doi.org/10.1145/2659522.2659531.

[3] Bhatti, S.K., Muneer, A., Lali, M.I., Gull, M. and Din, S.M.U. 2018. Personality analysis of the USA public using Twitter profile pictures. *2017 International Conference on Information and Communication Technologies, ICICT 2017*. 2017–Decem, (2018), 165–172. DOI:https://doi.org/10.1109/ICICT.2017.8320184.

[4] Big Five personality traits: *https://en.wikipedia.org/w/index.php?title=Big_Five_personality_traits&oldid=875501560*. Accessed: 2018-10-04.

[5] Carducci, G., Rizzo, G., Monti, D., Palumbo, E. and Morisio, M. 2018. TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning. *Information (Switzerland)*. 9, 5 (2018), 1–20. DOI:https://doi.org/10.3390/info9050127.

[6] Celli, F., Lepri, B., Biel, J.-I., Gatica-Perez, D., Riccardi, G. and Pianesi, F. 2014. The Workshop on Computational Personality Recognition 2014. *Proceedings of the ACM International Conference on Multimedia - MM '14*. (2014), 1245–1246. DOI:https://doi.org/10.1145/2647868.2647870.

[7] Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.F. and De Cock, M. 2016. Computational personality recognition in social media. *USER MODEL USER-ADAP*. 26, 2–3 (2016), 109–142. DOI:https://doi.org/10.1007/s11257-016-9171-0.

[8] Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M. and Davalos, S. 2014. A Multivariate Regression Approach to Personality Impression Recognition of Vloggers. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition - WCPR '14*. 1, (2014), 1–6. DOI:https://doi.org/10.1145/2659522.2659526.

[9] Farnadi, G., Zoghbi, S., Moens, M. and Cock, M. De 2013. Recognising personality traits using facebook status updates. *AAAI Workshop - Technical Report*. (2013), 14–18.

[10] Majumder, N., Politécnico, I., Soujanya Poria, N., Gelbukh, A., Nacional, I.P. and Cambria, E. 2017. Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intell. Syst.* 32, 2 (2017), 74–79. DOI:https://doi.org/10.1109/MIS.2017.23.

[11] MyPersonality Project: *mypersonality.org/wiki/doku.php*. Accessed: 2018-02-23.

[12] Nie, D., Guan, Z., Hao, B., Bai, S. and Zhu, T. 2014. Predicting personality on social media with semi-supervised learning. *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*. 2, (2014), 219–244. DOI:https://doi.org/10.1109/WI-IAT.2014.93.

[13] Pedregosa, F., Weiss, R. and Brucher, M. 2011. Scikit-learn: Machine learning in Python. *J MACH LEARN RES*. 12, Oct (2011), 2825–2830. DOI:https://doi.org/10.1007/s13398-014-0173-7.2.

[14] Pennebaker, James W and King, L.A. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*. 77, 6 (1999), 1296.

[15] Rangel Pardo, Francisco Manuel and Celli, Fabio and Rosso, Paolo and Potthast, Martin and Stein, Benno and Daelemans, W. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*. (2015), 1–8. DOI:https://doi.org/1613-0073.

[16] Salem, Marwa S. and Ismail, Sally S. and Aref, M. 2019. Personality Traits for Egyptian Twitter Users Dataset. *Proceedings of the 2019 8th International Conference on Software and Information Engineering*. (2019), 206–211. DOI:https://doi.org/10.1145/3328833.3328851.

[17] Sarkar, C., Bhatia, S., Agarwal, A. and Li, J. 2014. Feature Analysis for Computational Personality Recognition Using YouTube Personality Data set. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition - WCPR '14*. (2014), 11–14. DOI:https://doi.org/10.1145/2659522.2659528.

[18] Shoukry, Amira and Rafea, A. 2012. Preprocessing Egyptian Dialect Tweets for Sentiment Mining. *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*. (2012), 47–56.

[19] Tadesse, M.M., Lin, H., Xu, B. and Yang, L. 2018. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. *IEEE Access*. 6, c (2018), 61959–61969. DOI:https://doi.org/10.1109/ACCESS.2018.2876502.

[20] Tandera, T., Hendro, Suhartono, D., Wongso, R. and Prasetio, Y.L. 2017. Personality Prediction System from Facebook Users. *PROCEDIA COMPUT SCI*. 116, (2017), 604–611. DOI:https://doi.org/10.1016/j.procs.2017.10.016.

[21] Tighe, E.P. and Cheng, C.K. 2018. Modeling Personality Traits of Filipino Twitter Users. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. (2018), 112–122. DOI:https://doi.org/10.18653/v1/W18-1115.

[22] Varshney, V., Varshney, A., Ahmad, T. and Khan, A.M. 2018. Recognising personality traits using social media. *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017*. (2018), 2876–2881. DOI:https://doi.org/10.1109/ICPCSI.2017.8392248.

[23] Verhoeven, B., Daelemans, W. and De Smedt, T. 2013. Ensemble Methods for Personality Recognition. *Proceedings of the Workshop on Computational Personality Recognition: Shared Task*. 13, L (2013), 1–4.

[24] Wan, D., Zhang, C., Wu, M. and An, Z. 2014. Personality prediction based on all characters of user social media information. *COMM COM INF SC*. 489, (2014), 220–230. DOI:https://doi.org/10.1007/978-3-662-45558-6.

[25] Whitty, M.T., Doodson, J., Creese, S. and Hodges, D. 2018. A picture tells a thousand words: What Facebook and Twitter images convey about our personality. *PERS INDIV DIFFER*. 133, (2018), 109–114. DOI:https://doi.org/10.1016/j.paid.2016.12.050.

[26] Xue, D., Hong, Z., Guo, S., Gao, L., Wu, L., Zheng, J. and Zhao, N. 2017. Personality recognition on social media with label distribution learning. *IEEE Access*. 5, (2017), 13478–13488. DOI:https://doi.org/10.1109/ACCESS.2017.2719018.

[27] Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X. and Sun, J. 2018. Deep learning-based personality recognition from text posts of online social networks. *APPL INTELL*. 48, 11 (2018), 4232–4246. DOI:https://doi.org/10.1007/s10489-018-1212-4.