IJICTR
International Journal of Information & Communication Technology Research

Volume 15- Number 1 – 2023 (45 -55)

Recognizing Personality Traits using Twitter & Facebook for Arabic Speaking Users in Lebanon

Mokhaiber Dandash
Electrical and Computer Engineering
Artificial intelligence and robotics group
Ph.D. Student, University of Tehran
Beirut, Lebanon
m.dandash@ut.ac.ir

Masoud Asadpour*
Electrical and Computer Engineering
Assistant Prof. Director of Social Networks Lab
University of Tehran
Tehran, Iran
asadpour@ut.ac.ir

Received: 5 May 2022 – Revised: 25 August 2022 - Accepted: 27 November 2022

Abstract—**Nowadays, Social media is heading toward personalization more and more. People express themselves and reveal their beliefs, interests, habits, and activities, simply giving a glimpse of their personality traits. The thing that pushed us toward further investigating the mutual relation between personality and social media, taking into consideration the shortage in covering such important topic, especially in rich morphological languages. In this paper, we work on the connection between usage of Arabic language on social outlets (mainly Facebook and Twitter) and personality traits. We indicate the personality traits of users based on the information extracted from their activities and the content of their posts/tweets in Social Networks. We use linguistic features, beside some other features like emoticons. We gathered personality data using Arabic personality test based on Myers-Briggs Type Indicator (MBTI), which contains Thinking, Feeling, Intuition, Introversion, Sensation, Extroversion, Perceiving and Judgement traits. We collected our dataset from 522 volunteers, who permitted us to crawl their tweets and posts in Twitter and Facebook. Analysis of this dataset proved that some linguistic features could be used to differentiate between different personality traits. We used and implemented Deep Learning, and BERT to reveal personality and create a model for this purpose. Up to our knowledge, this is the first work on detection of personality traits from social network's data in Arabic language.**

Keywords: Personality Detection; Social Networks; Arabic Language Processing; Linguistic Features.

Article type: Research Article

Publisher: ICT Research Institute

I. INTRODUCTION

We are living in a society dominated by Social Media; the influence of such facility can change the life style in society, flourish economy, or even make conflicts. To understand the social media, we should learn how it evolves in the community and analyze the data in it.

Social media is facing an exponential increase in the usage and dependency in all aspects of life. Nowadays, with the spread of reviews, ratings,

* Corresponding Author

recommendations and other forms of opinion expression, thanks to the arrival of the interactive web, social media analysis has become one of the essential research fields whose applications is clearly noticeable in a range of domains from politics, business, tourism, education, health, etc.

Given the significance of social media, many research works were dedicated to this research area. However, most of these studies have concentrated on English and other European languages. Very little studies have actually talked about Analysis in morphologically rich languages such as Arabic, given the increasing number of Arabic internet users, online content in Arabic is growing on a daily basis.

Twitter and Facebook reveal many aspects of user's life, either by the account profiles, or by the usage of those accounts, one of these user aspects is personality. The main question is whether social media profiles can contribute in identifying personality, so that we can relate the information revealed by the social media contributors to their personality traits.

To give answer there are several challenges to overcome, most notably:

- Arabic language complexity and ambiguity.
- Lack of annotated dataset for such topic.
- Getting the personality trait test for our case study.
- Finding most suitable ways to get the data from our case study participants.
- Existence of our participants on social media outlets in the first place.
- Having the permission to crawl their data after getting the personality.
- Creating ways to enhance participation and promote the study in such topics.
- Dealing with multi-class topic.
- Finding efficient way to analyze and produce best model possible.

To address these concerns, we first got the questionnaire used by Myers-Briggs Type Indicator (MBTI) for Arabic language. Then we used the questionnaire to collect a sample from about 1600 Arab natives; knowing that collecting personality data is a challenge because the tendency to answer such questionnaire is not very high. We used different promotion methods to emphasize participants to share results and spread the questionnaire.

After checking the contributor's answers to the standard personality questionnaire, we get a group of 388 users in Facebook and 134 users in twitter that fully answered the questions and had public accounts. We gathered their timeline from their accounts in Facebook and their last 3200 tweets from their Twitter accounts, extracted some features, and applied different machine learning methods. We show in this paper that it is possible to reveal some information about personality traits even with this small amount of data.

This paper is organized into four sections: Section 1 gives an introduction on the personality detection problem and difficulties of Arabic language processing. In Section 2 our research method is explained. Section 3 explains the Experimental Results and finally conclusion comes in Section 4.

II. BACKGROUND

Arabic language is one of the dominant languages worldwide, it's one of the six official languages of the United Nations (UN), and it's spoken by more than 400 million speakers. It has three main forms: Classical Arabic; which is the language of the Quran (Islam's Holy Book); Modern Standard Arabic (MSA) and dialectical Arabic. MSA is the most expressive Arabic language form that is used both in writing and formal speeches.

Dialectical Arabic refers to all spoken forms in daily life. These forms change from one Arab country to another one and sometimes from one region to another in the same country.

Arabic Language is written from right to left, and is deprived of upper or lower cases. Its alphabet includes 28 letters. In addition to these vocal segments, the Arabic handwriting uses pronunciation marks as short vowels. These are located either above or below the letters to provide the correct pronunciation and enrich the meaning of the word. The majority of MSA texts are written without short vowels. This is because native or skilled speakers do not need diacritical marks in order to know a given text. The lack of diacritical marks in many of texts creates a verbal ambiguity problem that challenges computational systems. For example the word شعر may mean (شِعْرٌ poetry), (شَعْرٌ hair) or (شَعَرَ to feel).

Arabic language has a very complex and rich morphology, a word in Arabic exposes several morphological aspects such as derived or Inflectional morphology

Derived morphology is the way of creating a new word depending on an existing word, e.g. in English the adjective ''daily" is derived from the noun ''day". For example, the three letters ''كتب" is a root that means ''write" If it is put in the pattern (kataba, كَتَبَ), Or means Books If it is put in the pattern (كُتُب ,Kutub).

Inflectional morphology defines the deviation of a word to describe similar meaning in different grammatical groups (e.g. in English: eat, ate, eaten). The set of these changed word-forms called a lexeme class. To express the lexeme, a lemma, which is a particular form, is usually selected.

In Arabic, words change to different categories like tense (past and present), person (1st, 2nd and 3rd), number (singular, dual and plural), gender (feminine and masculine). For example, the inflection of the verb

‘‘كتب” (write) depending on past tense and first person, singular and masculine is “كَتَبْتُ”.

Arabic is an agglutinative language, which means that the word may be joined with a set of clitics (affixes). The English phrase ‘‘and with his life”, for example, corresponds to the Arabic form ‘‘وبحياته”. This word can be split into four parts (و + ب + حياة + ه) Multiple word prefixing, suffixing and affixing produce different word from the same stem.

The complexity of the Arabic word structure is one of the main difficulties that researchers face when dealing with Arabic Analysis.

A. Difficulties in Arabic Processing

The main purpose of morphological analysis is to divide words into morphemes and to tie up each morpheme with a morphological information such as stem, root, Part of Speech (POS), and affix. As we discussed before, Arabic is a morphologically complex language. This complexity needs the development of suitable systems that are able to work with tokenization, stemming, lemmatization, and POS tagging.

Nowadays, many morphological analyzers for Arabic are already developed; some of these are freely available while the rest have a profitable purpose. However, these systems suffer from important limitations especially in handling ambiguity that can result from the elimination of diacritics (vowels), or the free word-order nature of Arabic sentence.

For communication purposes, Arabic speakers usually use informal Arabic rather than MSA. There are around 30 major Arabic dialects that vary from MSA and from each other phonologically, morphologically, and lexically (Habash, 2010) [1]. Furthermore, Arabic dialects have no standard orthographies and no language colleges.

Therefore, using tools and resources designed for MSA to process Arabic dialects produces clearly low performance. Recently, researchers have started developing parsers for specific dialects such as CALIMA (Habash et al., 2012) [2] for Egyptian Dialect. However, these analyzers still have low accuracy, and is used only for certain dialects. Filling this gap in processing Arabic will enhance effectiveness of information retrieval especially for social media data.

Abdul-Mageed and Diab offered a large-scale sentiment lexicon (Abdul-Mageed and Diab, 2014) [3]. This lexicon consists of entries covering MSA and several Arabic dialects. They gathered two-word lists from Penn Arabic Treebank and Yahoo and labelled them manually. Regardless of the large size of the created resource, many of the entries are not lemmatized which restricts the usage of their lexicon.

In their effort to form Arabic multi–domain resources, ElSahar and El-Beltagy offered a semi-supervised approach to produce domain lexica out of the four reviewed datasets (ElSahar and El-Beltagy, 2015) [4]. This method uses the feature selection abilities of SVM to select the most effective unigram and bigram features. Although the created lexicon contains diverse domains, it is pulled out only from reviews, which limits its application in social media analysis only to Sentiment Analysis.

Badaro et. Al. (Badaro et al., 2014) [5] created ArSenL, a lexicon for Arabic sentiments via two approaches depending on English SentiWordNet (ESWN). The first method associates each term in ArabicWordNet with ESWN to have sentiment grade, and with SAMA (Standard Arabic Morphological Analyzer) to discover the right lemma formulas. In the second approach, English annotations accompanying with SAMA’s entry are discovered automatically to find the most like synset in ESWN. The union of the two resulting lexica has a good exposure but is restricted to MSA.

In (M. Abdul-Mageed and M.T. Diab, 2012) [6], a corpus of MSA was gathered from different forums. One portion of the corpus was tagged using crowdsourcing on Amazon Mechanical Turk. Students who got exact instructions did tagging of the second portion. The last section annotated by students but using simple rules. However, this corpus is not publicly accessible and not used in any other work.

In addition to these corpora, an Arabic Twitter corpus was gathered in (E. Refaee and V. Rieser, 2014) [7] by Twitter API and checked in a pre-processing step. Two Arab natives tagged manually about 9000 Tweets using four labels: *neutral*, *mixed*, *positive* and *negative*. Morphological, syntactic, and semantic features were also added to the annotation.

B. Personality Test

There is nothing significant done for personality analysis in Arabic language from social networks. However, there are plenty of works done for English language such as [8-15]. They mainly concentrate on the Myers-Briggs Type Indicator (MBTI) or Big Five (BF) traits, and rely on “My Personality” (a Facebook application used to reveal personality based on an online questionnaire) dataset. There are other works, which do not depend on social network’s textual data but on the pictures [10] and on other engagements features.

Figure 1. Big Five Personality Factors [16]

The Big Five (BF) personality traits (Fig. 1), also known as the five-factor model (FFM) and the OCEAN model, is a classification for personality traits. The five factors are:

-Openness to experience: reflect degree of curiosity, creativity and novelty. People in this category are independent and open-minded.

-Conscientiousness: people in this category are organized and dependable, and prefer planned rather than spontaneous behavior.

-Extraversion: people in this category are energetic, attention seeking, talkative, and tend to expresses their ideas.

-Agreeableness: They tend to be compassionate and cooperative rather than suspicious, and non-argumentative person.

-Neuroticism: They tend to be stressed, and to express unpleasant emotions like anger, anxiety.

The BF personality traits as mentioned in [17-21] are "the model to comprehend the relationship between personality and academic behaviors, These five overarching domains have been found to contain and include most known personality traits and are assumed to represent the basic structure behind all personality traits".

"Each of the Big Five personality traits contain two separate but correlated aspects reflecting a level of personality below the broad domains but above the many surface scales that are also part of the Big Five" [22].

The aspects are as follows:

- Volatility and Withdrawal for Neuroticism;
- Enthusiasm and Assertiveness for Extraversion;
- Intellect and Openness for Openness to Experience;
- Industriousness and Orderliness for Conscientiousness; and
- Compassion and Politeness for Agreeableness.

-According to [23] [24] [25] Myers–Briggs Type Indicator (MBTI) is *"An introspective self-report questionnaire indicating different psychological preferences in how people perceive the world around them and make decisions."*

-With the concept [26]: *"MBTI has speculated that humans experience the world using four principal psychological functions – sensation, intuition, feeling, and thinking – and that one of these four functions is dominant for a person most of the time".*

The MBTI sorts some of these psychological differences into four opposite pairs, with a resulting 16 possible psychological types.

The opposite four pairs are:

1. Extraversion (E) vs. Introversion (I)
2. Sensing (S) vs. Intuition (N)
3. Thinking (T) vs. Feeling (F)
4. Judgment (J) vs. Perception (P)

The 16 types typically referred to by an abbreviation of four letter i.e. the initial letters of each of their four type preferences (except in the case of intuition, which uses the abbreviation "N" to distinguish it from introversion). For instance:

- ESTJ means Extraversion (E), Sensing (S), Thinking (T), Judgment (J)
- INFP means Introversion (I), Intuition (N), Feeling (F), Perception (P)

In overall, personality detection through social media has not yet been a subject of research in Arabic; however, there are many works in English getting benefit of twitter GNIP API, and "MyPersonality" Facebook dataset.

III. METHODOLOGY

Social media analysis is not a standalone concept, from our perspective; it should have assisting algorithms to extract knowledge from it in a human-understandable structure. Assisting algorithms that we are trying to enhance by the work of this paper.

To reach high levels in data analysis challenges will come across, such as the way of collecting data, analysis standards, reaching a good interpretation of data, deciding what kind of data to use from social media outlets, just to mention a few.

Social media outlets give us a way to interact with other users, (like/share their work, post video, links, etc.) even if long distances separate us. It makes a forum for people to know each other, and it is simply the context of social media. Users' personality is one of the main reasons why the content of the posts and interactions on social media differ from one user to another. Let us wonder what contribution personality has on influencing social media data, numbers of likes, number of sharing, using retweets, or the abundance of posts and tweets.

In personality analysis for Arabic language, there is no existing dataset. After the Cambridge Analytica crisis [27] that used MyPersonality App (Fig. 2) to collect personality data from Facebook accounts, collecting personality-related datasets has been even more difficult.

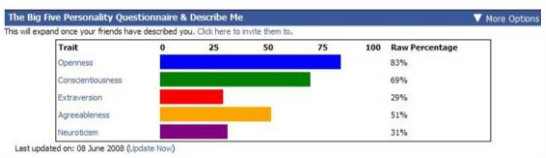


Figure 2. My personality App sample from facebook [11]

To solve this dilemma, we worked along with psychology professors in the Lebanese university and

institutes in Lebanon that work in human development projects. We got the personality questionnaire that is specialized for Arabic speakers that is tested and oriented toward university students [28], in order to identify what faculty suit the students more. Then we created three versions of the questionnaire: an offline version (printed papers), and two online versions: google forms and a website (Fig. 3).

In this research like other similar researches, we followed the following steps:

1. We first ask some users to participate in the standard personality test, executed on the google forms, our designed website or the questionnaire printed on the paper.
2. They participate in the test and their score is calculated.
3. The result of the test will be sent to them as a feedback via email or a web page (Fig. 4).
4. They are asked to reveal their account in Facebook and Twitter and allow us to collect their activities in these platforms.
5. If they accept and the accounts are not private, we crawl their accounts and record the necessary information that is accessible from their activities (posts, tweets, etc.).
6. The recorded information is converted into a vector composed of quantitative and categorical features using different methods.
7. Different Machine learning techniques are applied to relate the features that come from social media to personality scores.
8. Once the mapping technique is developed, we are able to assign personality scores to other people, which have not participated yet in the personality assessment test but their account in social media is accessible.

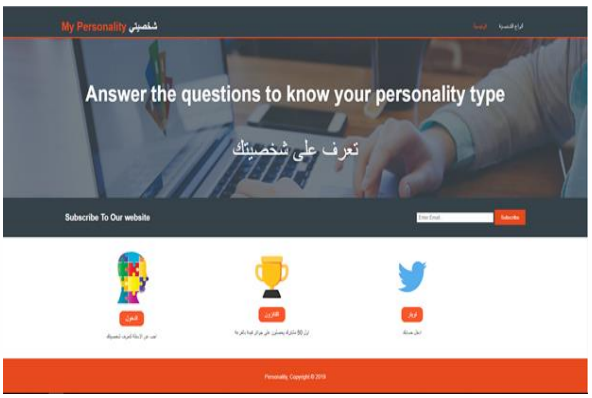


Figure 3. Web site designed for running the personality test

Figure 4. Personality features explained in the result

IV. EXPERIMENTAL RESULTS

Users interact with each other more, when more similarities between their personalities exist, “Similarity breeds connection” [29]. We get relations between users by actions done on their social media accounts (e.g. retweet/share).

We gathered information from social media outlets, we chose Facebook and Twitter and we collect data using their official API, as shown in the figure below:

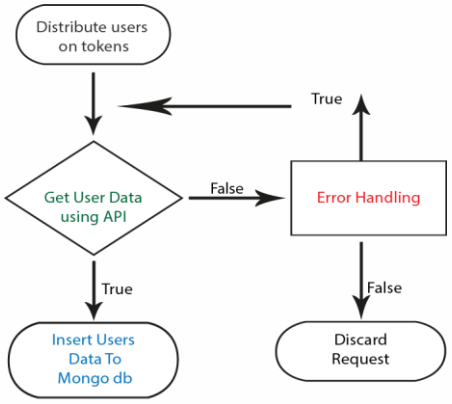


Figure 5. Data gathering flow chart

All the experiments implemented in Python with Desktop computer. The processor is Intel® Core™ i7-7700 CPU @ 3.60 GHz and 16GB of RAM. GPU is Nvidia GeForce GTX 1060 6GB

A. Subjects of the experiment

The people that participated in our test were mainly university students (students of Lebanese university and other universities like ALMAAREF and LIU university), and youth of different genders (68% female), religions, education and regions in Lebanon. Later when we used Google forms and website, participants from other age categories participated in our work (We invited people to use google forms through sending the link to university groups, and sharing it across members and their families).

Gathering the needed dataset for personality was the most important step, because the dataset is the fuel of our work. We first printed 2000 form of the personality test, and spread it across different departments (medicine, science, journalism and engineering) in the Lebanese universities; this gave us more diversity of participants. This way we gathered about 600 answers, from which about 280 were useful,

and from those 280 only 125 user had public accounts and have spoken before on social media and we could verify their accounts.

Then, we tried the second method to get online data through google forms. We spread promotions to enhance the participation of people via filling google forms and got about 926 contributions, ensuring that the test results were sent through email to the participants immediately.

We also built a website to get the personalities of people using the "continue to Facebook" button, beside the option to add social accounts and we got some contributors as well.

Therefore, as total we got about 1600 contributor, among them the data of 522 persons were useful, from which 388 were from Facebook and 134 from twitter, the statistic that is significant compared to works done before with 50 users [9]. We got a dataset of 117862 text distributed over 16 personality trait, the gathered data transferred in JSON form to MongoDB for easier handling.

B. Extracted Features

Since the collected records from Facebook and Twitter have different fields, we trained different models. Some features are common between both models but some features are unique to just one of them. It is important to mention that after collecting the data we made a normalization step so that we removed the stop words and all characters other than Arabic letters.

We have used the following features:

1) Bag of words (Sivic et al., 2009) [30]: we denoted the existence of informative words (excluded some useless stop-words from a list) by assigning a column to that word and coded the column corresponding to that word with one (1) and set the other entries to zero (0). For every text, the entries of this vector has one in the columns that corresponds to the words that exist in the text and has zero for the other entries.
2) Term Frequency Inverse Document Frequency (TF-IDF) (Rajaraman et al., 2011) [31]: This is the same as the previous vector except that instead of ones, the TF*IDF of the word (which is a real number) is used.
3) Word2vec (Mikolov et al., 2013) [32]: The texts are given to a Neural Network that transfers the words into a 300-axes space to use in Deep Learning. The network encodes the proximity of the words that appear near to each other in the text and for each word it gives a vector that uniquely specifies it in that space.
4) Multi-linguality: Some contributors had posts in English, the thing that we did not like first, because our goal is Arabic language, yet we used this feature as an evidence to strengthen the results of some personality types. Some users may like to use foreign language for e.g. to show off, which might reveal some part of their personality traits.
5) Number of Likes: the number of likes by the user is a sign for his interests in different activities, and his way to express more feelings.
6) Past actions: Many researchers say that knowing a person, can be done by checking his past actions, because action just shout loudly the personality traits of each person. Having the social media profile of a person is just a brief of his past, which can be as a feature. In the paper by (F. Celli, 2012) [33] eight persons just predict the personality of 142 twitter user after checking their tweets.
7) Gender (categorical)
8) In (Quercia et al., 2011) [11], and (Golbeck et al., 2011) [34] both papers used features from LIWC (LINGUISTIC INQUIRY AND WORD COUNT) to check influence on personality. we used here number of hashtags
9) Number of words used by users of every personality trait
10) emoji usage and number of usages (by counting number of emoji's used by users of every personality trait)
11) Word density (number of characters divided by word count)
12) Number of punctuation characters
13) Number of words of the title (count the word start with an upper case letter)
14) Number of characters
15) Relation between language features and personality traits (phrases and expressions used frequently that reflect features for specific personality like: first person, compassion, love, opinion, happiness, and sadness expressions) explained in next section.

C. *Results*

We used different information extraction methods (like bag of words and TF-IDF) in order to find any relation between text features and personality traits. We also used Pearson correlation coefficient denoted by r to find the relation between the personality and text features. Pearson correlation as defined by Wikipedia: is a statistic that measures the linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation. 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.

The tests showed that people with INFP (the healer) and INTP (the counsellor) personalities have used the maximum number of emoji. This show in a good manner that introversion, intuition and perception people use more emoji to express themselves.

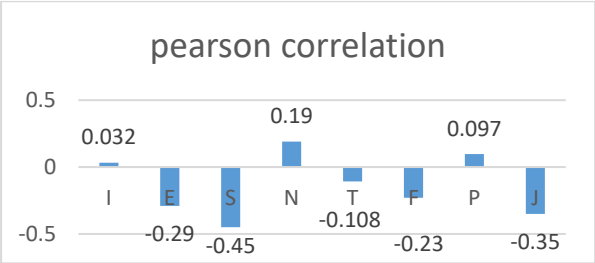


Figure 6. Pearson correlation with respect to personality trait

We found that the Pearson coefficient (figure6) for the following personality trait introvert, intuition and perception are positive, this means that when the number of words increase the number of emoji's increase. The thing that make sense because introversion people tend to use some kind of symbols to state their situation, also for perception and intuition is same as people with perception go with the flow before intuitive thinking.

Results shown below are demonstrating some features for each personality traits; we can see that in the case of word count, punctuation count, title word count and character count (figure 7,9,10,11) the Feeling and Sensing people wrote more than the others and are more like to use details.

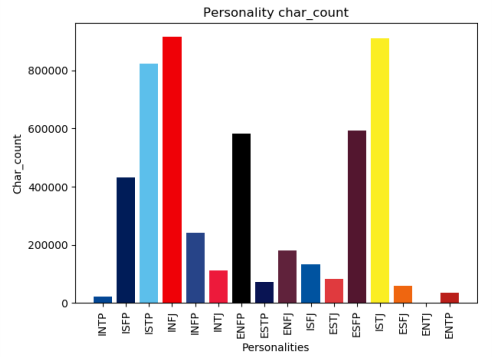


Figure 7. Character count with respect to personalities

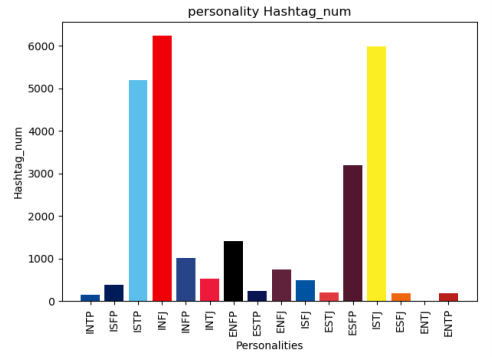


Figure 8. Hashtags number with respect to personalities

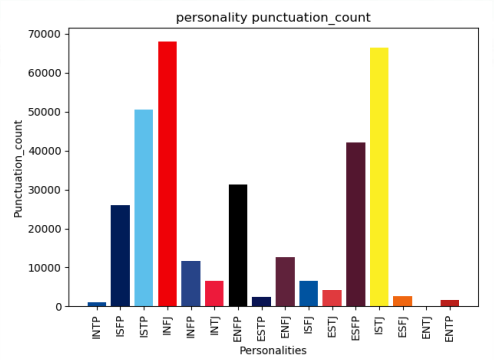


Figure 9. Punctuation count with respect to personalities

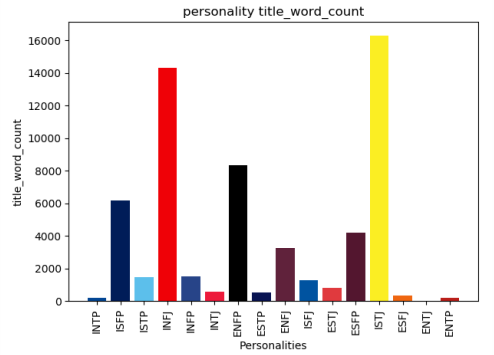


Figure 10. Big Title word count with respect to personalities

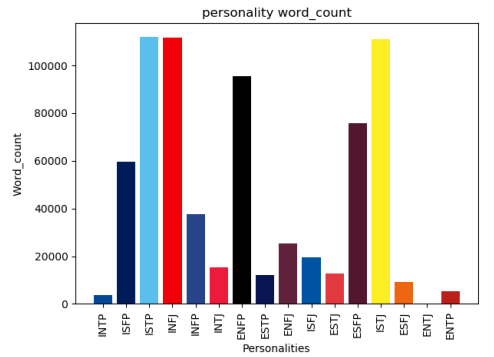


Figure 11. Word count with respect to personalities

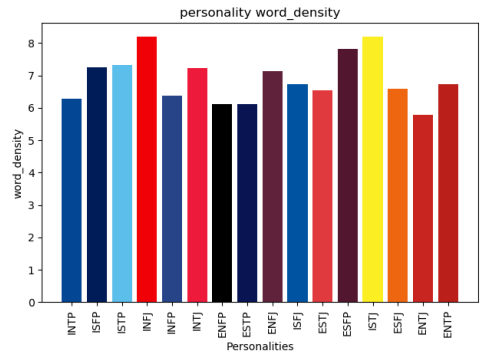


Figure 12. Word density with respect to personalities

Fig 8 shows that INFJ (the Counsellor), ISTJ (the inspector), ISTP (the Craftsman) are the personality traits who have used hashtags the most. This shows that

sensing and feeling people use many hashtags. The other note is that the introversion people, in overall, use more hashtags than extroversion people do. Maybe because they do not like to write too much text, instead they express themselves in hashtags.

We used the most frequent 1000 unigram, bigram, trigrams in our dataset. The purpose is to effectively, decrease the size of feature sets while still keeping the valuable indications in these linguistic features. The bag of words used to measure the Correlation between unigram words and personality traits, for example:

Figure 13. Correlated unigram ESTJ

Figure 14. Correlated unigram for ISFJ

As we can see in the case of thinking and extroversion specific words are used like (serious, جدي) and (افضل, I prefer) in case of feeling and introversion we have more soft words like (عاطفيا, الحياة), (life, emotional) this may give a good hint on the relation between phrases usage and personality traits.

Now we discuss our model that aim to predict personality depending on the features (bag of words, TF-IDF, TF-IDF n-gram, TF-IDF with n-gram on character base, and word vector), we develop different prediction models using both individual features and combined features. For combined features, we concatenate features within to test the predictive power of our models. First, we used different machine learning algorithms as shown in Fig 15:

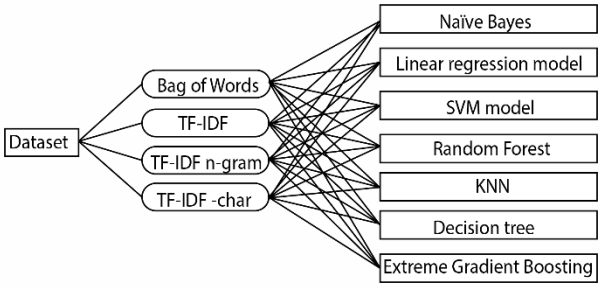


Figure 15. Applied machine learning algorithms

TABLE I. THE PERFORMANCE OF DIFFERENT MACHINE LEARNING METHODS WITH BAG OF WORDS FEATURE

Bag of words	Accuracy	f-measure
Naïve Bayes	26.78%	14.94%
Linear regression model	29.38%	16.24%
SVM model	19.06%	3.21%
Random Forest	34.64%	23.59%
KNN	27.66%	13.06%
Decision tree	34.74%	24.15%
Extreme Gradient Boosting	27.21%	17.56%

TABLE II. PERFORMANCE OF DIFFERENT MACHINE LEARNING METHODS WITH TF-IDF FEATURE.

TF-IDF	Accuracy	f-measure
Naïve Bayes	26.04%	11.74%
Linear regression model	29.85%	13.79%
SVM model	17.05%	1.82%
Random Forest	35.04%	24.67%
KNN	27.98%	12.98%
Decision tree	34.86%	23.55%
Extreme Gradient Boosting	27.37%	18.43%

TABLE III. PERFORMANCE OF DIFFERENT MACHINE LEARNING METHODS WITH BOTH TF-IDF AND N-GRAM (2, 3) FEATURE

TF-IDF n-gram	Accuracy	f-measure
Naïve Bayes	23.02%	10.25%
Linear regression model	24.00%	11.57%
SVM model	16.52%	1.77%
Random Forest	24.30%	13.04%
KNN	17.99%	10.24%
Decision tree	24.34%	13.16%
Extreme Gradient Boosting	22.55%	12.26%

TABLE IV. PERFORMANCE OF DIFFERENT MACHINE LEARNING METHODS WITH TF-IDF ON CHARACTER N-GRAMS

TF-IDF - char	Accuracy	f-measure
Naïve Bayes	29.97%	12.21%
Linear regression model	33.36%	14.58%
SVM model	17.11%	1.82%
Random Forest	40.58%	27.78%
KNN	18.98%	11.88%
Decision tree	39.38%	27.10%
Extreme Gradient Boosting	34.14%	20.59%

The result was not satisfying so we worked on feedforward neural network hoping in significant increase in accuracy, the results we got as below:

TABLE V. NEURAL NETWORK RESULT WITH DIFFERENT FEATURE EXTRACTION METHODS

	Neural Network
Bag of words	33.82%
TF-IDF	34.15%
TF-IDF (n-gram(2,3))	23.98%
TF-IDF(n-gram character base)	39.35%

Still not convincing results, so we suggested feedforward deep learning and (LSTM) method that combine between word vector and bag of words feature extraction, and we got the result of 54.5% as the highest accuracy.

TABLE VI. DEEP LEARNING RESULT WITH DIFFERENT FEATURE EXTRACTION METHODS

	LSTM	Deep learning
Bag of words and word2vec	27.30%	54.5%
TF-IDF and word2vec	18.34%	22.58%
TF-IDF (n-gram(2,3)) and word2vec	19.78%	20.47%

TF-IDF(n-gram character base) and word2vec	17.12%%	16.87%

The practice showed that using combined features improve the model performance, removing word2vec shows significant drop in accuracy.

At last, in order to concentrate more on the context of the words, knowing that personality traits can be hidden in context, we worked on state-of-the-art method BERT [35]. Deep learning based NLP models show key enhancements when trained on millions of annotated training samples, the thing that is not always available, to link the gap in data, researchers worked on creating methods for training general-purpose language illustration models using the unannotated text (pre-training). These general-purpose pre-trained models can then be fine-tuned on smaller task-specific datasets.

Pre-trained language illustrations can be either context-free or context-based. Context-based models generate a representation of each word based on the other words in the sentence, the thing that will help our work especially with complexity of Arabic language.

BERT is based on the Transformer model architecture, instead of LSTMs. Transformer (the attention mechanism that learns contextual relationships between words in a text) in general consists of an encoder to read the text input and a decoder to produce a prediction for the task. Since BERT's objective is to produce a language representation model, it only wants the encoder part. The input to the encoder for BERT is a sequence of tokens, which are first changed into vectors, processed in the neural network. However, before processing can start, BERT needs the input to be changed and decorated with some extra metadata.

A Transformer works by performing a small, constant number of steps. In each step, it applies an attention mechanism to understand relationships between all words in a sentence, regardless of their respective position.

Overall process divided into two steps:

-pre-train:
1. Masked LM (MLM)
2. Next Sentence Prediction (NSP)

-fine tune: how to use language for specific task

We can see that BERT can be applied to many diverse tasks by adding a task-particular layer on top of pre-trained BERT layer. For text classification, we can just add the classifier to the top of BERT, as shown in Fig 15.

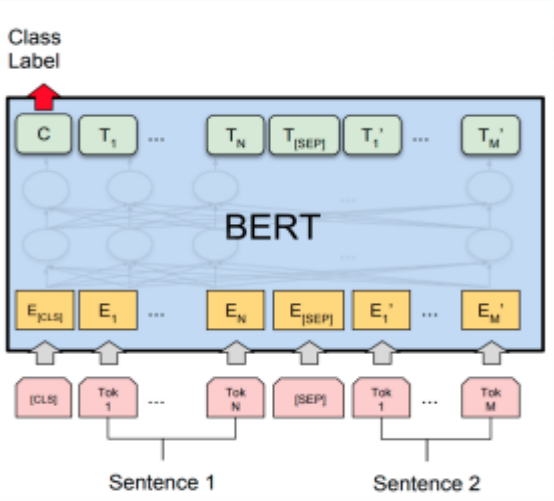
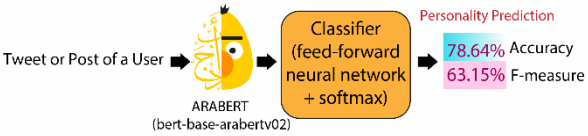


Figure 16. BERT fine-tuning [35]

We used general purpose pre-trained BERT for Arabic AraBERT [36] with "bert-base-arabertv02" as model, prepared the data by changing personality label into representative label, we apply pre-processing using AraBERT processor. We initialized the pre-trained model and tokenizer, we check the tokenized sentence length to decide the maximum sentence value length, we make pre-trained model ready for classification and used one classification layer with evaluation metrics F-measure and accuracy. Our training loop is 16 epochs, learning rate=2e-5, and batch size=16.



As shown in the figure above, fine-tuning this pre-trained language model with our dataset got an improvement in our accuracy result 78.64%, outperformed other methods.

V. CONCLUSION

In this paper, we analyzed the collected data from the Facebook and Twitter accounts of some Arabic speaking users and using the results of their personality tests, we extracted some features that helped us detect their personality traits with about 79% accuracy. The best results gained by using Arabic BERT model.

The model available now, can be used for understanding the personality of other Arabic-speaking social media users. This work is an ongoing research and we are collecting more profiles in order to increase its accuracy and F-measure.

This model can then be used in applications e.g. in Recommender systems, Customer relation management (CRM), Attracting new clients, retaining current clients, bring back former clients, reduce marketing costs, and Identifying purchasing habits. Also targeting people in election or influencing them may benefit from such model.

To conclude, Arabic social analysis with all its different fields is necessity to work on, especially with the exponentially increasing users of Arabic language on the web. Every aspect added will make the analysis narrow more, to get exact orientation of the text in social media, the thing that we delivered in this paper with the models we got.

In the future, we are willing on improving these results by increasing the data set. We will work on a new novel way of collecting personality results to ensure diversity of data, with bigger spectrum of features to analyze.

REFERENCES

[1] Habash NY. Introduction to Arabic natural language processing. Synth. Lect. Hum. Lang. Technol. 2010;3(1):1–187.

[2] N. Habash, R. Eskander, A. Hawwari, A morphological analyzer for Egyptian Arabic, Presented at the Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, 2012, pp. 1–9.

[3] M. Abdul-Mageed, M.T. Diab, SANA: a Large Scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis, Presented at the LREC, 2014, pp. 1162–1169.

[4] ElSahar H, El-Beltagy SR. Building large Arabic multi-domain resources for sentiment analysis. In: Computational Linguistics and Intelligent Text Processing. Springer; 2015. p. 23–34.

[5] Badaro G, Baly R, Hajj H, Habash N, El-Hajj W. A Large Scale Arabic sentiment lexicon for Arabic opinion mining. ANLP 2014.

[6] M. Abdul-Mageed, M.T. Diab, AWATIF: a multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis, Presented at the LREC, 2012, pp. 3907–3914.

[7] E. Refaee, V. Rieser, An Arabic twitter corpus for subjectivity and sentiment analysis, Presented at the LREC, 2014, pp. 2268–2273.

[8] Yilun Wang, Understanding Personality through Social Media Department of Computer Science, Stanford University.

[9] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (Social-Com), pages 149–156. IEEE.

[10] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E. Moghaddam, Lyle Ungar, Analyzing Personality through Social Media Profile Picture Choice.

[11] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, “Our Twitter Profiles, Our Selves: Predicting Personality with Twitter,” in *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, 2011, pp. 180–185.

[12] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. 2012. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. (ICMLA), 2012 11th International Conference on Machine Learning and Applications, volume 2, pages 386–393. IEEE.

[13] H.Andrew Schwartz, Johannes C.Eichstaedt ,Lukasz Dziurzynski,, Margaret L.Kern, Martin E.P.Seligman, Lyle H.Ungar, Eduardo Blanco, Michal Kosinski and David Stillwell, Toward Personality Insights from Language Exploration in Social Media. 2013 AAAI Spring Symposium Series.

[14] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, Martine De Cock, Recognizing Personality Traits Using Facebook Status Updates. Seventh International AAAI Conference on Weblogs and Social Media.2013

[15] Fabio Celli, Fabio Pianesi, David Stillwell, Michal Kosinski. Workshop on Computational Personality Recognition: Shared Task. Seventh International AAAI Conference on Weblogs and Social Media.2013

[16] Big Five - the personality in five dimensions https://peats.de/article/big-five-die-personlichkeit-in-funf-dimensionen

[17] *Shrout PE, Fiske ST (1995). Personality research, methods, and theory. Psychology Press.*

[18] Allport GW, Odbert HS (1936). "Trait names: A psycholexical study". PsychologicalMonographs. **47**:211. doi:10.1037/h0093360

[19] Bagby RM, Marshall MB, Georgiades S (February 2005). "Dimensional personality traits and the prediction of DSM-IV personality disorder symptom counts in a nonclinical sample". Journal of Personality Disorders. **19** (1): 53–67. doi:10.1521/pedi.19.1.53.62180. PMID 15899720

[20] Tupes EC, Christal RE (1961). "Recurrent personality factors based on trait ratings". USAF ASD Tech. Rep. **60** (61–97): 225–51. doi:10.1111/j.1467-6494.1992.tb00973.x. PMID 1635043.

[21] Norman WT (June 1963). "Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings". Journal of Abnormal and Social Psychology. **66** (6): 574–83. doi:10.1037/h0040291. PMID 13938947

[22] DeYoung CG, Quilty LC, Peterson JB (November 2007). "Between facets and domains: 10 aspects of the Big Five". Journal of Personality and Social Psychology. 93 (5): 880–96. doi:10.1037/0022-3514.93.5.880. PMID 17983306.

[23] Myers, Isabel B.; Myers, Peter B. (1995) [1980]. Gifts Differing: Understanding Personality Type. Mountain View, CA: Davies-Black Publishing. ISBN 978-0-89106-074-1.

[24] "MBTI® Basics". The Myers & Briggs Foundation. Archived from the original on 2021-10-12. Retrieved 2021-10-28.

[25] "Myers-Briggs Type Indicator® (MBTI®) | Official Myers Briggs Personality Test". www.themyersbriggs.com.

[26] Huber, Daniel; Kaufmann, Heiner; Steinmann, Martin (2017). The Missing Link: The Innovation Gap. Bridging the Innovation Gap. Cham: Springer International Publishing. pp. 21–41. doi:10.1007/978-3-319-55498-3_3. ISBN 978-3-319-55497-6. Retrieved 2021-10-28.

[27] Jim Isaak, Mina J.Hana (2018). "User data privacy: Facebook,Cambridge Analytica, and privacy protection". Ieee.org computer 51 (8),56-59.

[28] http://jwm.life/uploads/mbti/mbti-test-eman-azmi1.pdf

[29] M. McPherson, L. Smith-Lovin, and J. Cook. (August 2001). “Birds of a Feather: Homophily in Social Networks” Annual **Review of Sociology;** Vol. 27: 415-444.

[30] Sivic, Josef (April 2009). "Efficient visual search of videos cast as text retrieval". IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4. IEEE. pp. 591–605.

[31] Rajaraman, A.; Ullman, J.D. (2011). "Data Mining". Mining of Massive Datasets. pp. 1–17.

[32] Tomas Mikolov; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". International Conference on Learning Representations (ICLR 2013).

[33] F. Celli, “Unsupervised Personality Recognition for Social Network Sites,” in *ICDS 2012, The Sixth International Conference on Digital Society*, 2012, no. c, pp. 59–62.

[34] J. Golbeck, C. Robles, and K. Turner, “Predicting personality with social media,” in *CHI’11 extended abstracts on human factors in computing systems*, 2011, pp. 253–262.

[35] *Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".* arXiv:1810.04805v2.

[36] @inproceedings{antoun2020arabert,title={AraBERT: Transformer-based Model for Arabic Language Understanding},author={Antoun, Wissam and Baly, Fady and Hajj, Hazem},booktitle={LREC 2020 Workshop Language Resources and Evaluation Conference 11--16 May 2020},pages={9}}.

Masoud Asadpour received his B.Sc. degree in Computer Software Engineering from Sharif University of Technology, Tehran, Iran, in 1997, M.Sc. degree in Machine Intelligence and Robotics from University of Tehran, Iran, in 2000, and the Ph.D degree in Robotics from EPFL University, Lausanne, Switzerland, in 2007. He is an Assistant Professor with the University of Tehran, Director of Social Networks Lab. His current research interests are Social Network Analysis and Mining, Big Data Processing, Natural Language Processing and Machine Learning.

Mokhaiber Dandash recieved his B.Sc. degree in Electrical Control Engineering from Shahed University, Tehran, Iran, in 2013, and M.Sc. degree in Industrial Automation from Tarbiat Modares University, Tehran, Iran, in 2016. He recieved his Ph.D. in Artificial Intelligence and Robotics from University of Tehran, Iran in 2016. His current research interests are Social Network Analysis and Mining, Big Data Processing, Natural Language Processing and Deep Learning.