



Personality Analysis Using Classification on Turkish Tweets

Gokalp Mavis, Middle East Technical University, Turkey

Ismail Hakki Toroslu, Middle East Technical University, Turkey

 <https://orcid.org/0000-0002-4524-8232>

Pinar Karagoz, Middle East Technical University, Turkey

 <https://orcid.org/0000-0003-1366-8395>

ABSTRACT

According to the psychology literature, there is a strong correlation between personality traits and the linguistic behavior of people. Due to the increase in computer based communication, individuals express their personalities in written forms on social media. Hence, social media has become a convenient resource to analyze the relationship between personality traits and linguistic behaviour. Although there is a vast amount of studies on social media, only a small number of them focus on personality prediction. In this work, the authors aim to model the relationship between the social media messages of individuals and big five personality traits as a supervised learning problem. They use Twitter posts and user statistics for analysis. They investigate various approaches for user profile representation, explore several supervised learning techniques, and present comparative analysis results. The results confirm the findings of psychology literature, and they show that computational analysis of tweets using supervised learning methods can be used to determine the personality of individuals.

KEYWORDS

Big Five Personality Traits, Deep Learning, Machine Learning, Personality, Social Media, Supervised Learning, Tweets

1. INTRODUCTION

Personality is one of the typical and enduring topics in psychology. Personality prediction can be basically defined as identifying personality traits of a person by using a set of data. Current personality studies often rely on data from well-controlled and specified environments (Rozin, 2001), such as survey studies. However, social network usage is getting more and more popular and the data produced by social media users can provide a valuable resource to automatically determine human personality. Social media is one of the most commonly used services on the Internet. Studies show that one third of the time spent on Internet is allocated on social media sites (Suhartono et al., 2017). Although the use of social networks is increasing day by day, the number of studies focusing on the relation of social media and personality is still limited, and there are open problems to be studied such as analyzing the effective and language dependent features and the use of deep neural models for the problem (Ahmad, 2020) (Bharadwaj, 2018).

DOI: 10.4018/IJCINI.287596

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Psychologists show that there is a strong relationship between the personality and the language behaviour of individuals (Fast & Funder, 2008). Nowadays, computer based written communications became as common as face to face communication. Therefore, there are vast number of research efforts to determine personality traits of individuals from the texts they have written. Among the mediums, social media is probably by far the most popular venue people use. However, people write very informally and short texts in those places. Therefore, in this work we focus on such kind of texts and aim to determine the personality traits of individuals from their short informal texts.

Most of the previous studies about personality analysis are conducted on samples provided under controlled conditions, such as talks or texts on given topic (Baddeley & Singer, 2009), (Fast & Funder, 2008), (Hirsh & Peterson, 2009), (Pennebaker & King, 2000). On the other hand, naturalistic approach is proven to be more powerful in (Mehl, Gosling, & Pennebaker, 2006), where samples of participants' natural language use and behavior are recorded and analyzed. The study in (Mehl, Gosling, & Pennebaker, 2006) provides useful results about the relation between natural language use and personality, most of which were not captured in laboratory studies.

Identifying users' personality can be useful in different domains such as customer analysis for commerce, social psychology or recommendation systems to build personalized models and to improve user experiences. Advertising can be another field that can benefit from personality analysis. In (Odekerken, De Wulf, & Schumacher, 2003), the relation between consumer personality and marketing techniques is presented. In the last decades, psychologists who work on personality analysis by using lexical approach, ended up viewing personality in five dimensions (Boele, 2000). This approach is known as *Big Five personality traits* (Goldberg, 1990). This model alleges that dimensions of *neuroticism*, *openness*, *extraversion*, *agreeableness* and *conscientiousness* can comprise most of the structure of the personality traits. Big Five personality traits are known as the most widely accepted personality dimensions in psychology (Zhang, 2002).

Extraversion (also known as surgency), is the degree for dimensions like activeness, pretentiousness, talkativeness (Ashton, Lee, & Paunonen, 2002). *Neuroticism* describes emotional stability of people according to level of anxiety, depression and nerve (Widiger & Cost, 2012). *Agreeableness* is a dimension to describe characteristics like gentleness and kindness (Graziano & Eisenberg, 1997). *Conscientiousness* is about being respectful, organized, and trustworthy (Lepine, Colquitt, & Erez, 2006). *Openness* basically tells about openness to experience new things. This dimension covers personality attributes like creativeness and introspectiveness (McCrae & Costa, 1997).

In this work, we study the problem of personality trait prediction through text. Rather than using formally written texts on a given topic, we follow a naturalistic approach and use social media postings as the text resource to be analyzed. More specifically, Twitter posts (tweets) are collected as the data set. In addition to textual content, in the analysis, social media user statistics, such as the number of followers, followees and tweets posted, are considered as well. The proposed method includes the following steps:

- **Annotation:** The annotation of the data set that is needed to construct the supervised learning model is obtained through a conventional survey for Big Five personality traits. To this aim, we constructed a web based tool that presents the survey and matches the result with anonymized social media profile.
- **Natural Language Processing (NLP) and Text Mining:** NLP and text mining methods are used in order to analyze text content in social media platforms. We worked on various feature extraction methods and applied feature filtering. This step constructs the social media profiles based on the constructed and the selected features.
- **Classification:** Supervised machine learning models are constructed by using annotated social media user profiles as the training data.

Although both are based on Roman letters, Turkish alphabet is slightly different from English alphabet. In Turkish alphabet there are 6 additional letters and it lacks three letters of English alphabet (Q, X, W). However, since people write very informally in tweets, sometimes they also use English alphabet to write Turkish tweets (such as typing *s* instead of *ş*, or *u* instead of *ü*) and thus this also adds another form of misspelling. However, from the text and word's structure, the correct form can be constructed in most cases. There are also some cultural differences which effect the forms of misspellings preferred by the tweet writers. In this study our focus is on Turkish informal texts, which incurs additional difficulties due to the use and the structure of the language. To the best of our knowledge, there is no previous research in the literature specifically using informal Turkish texts in order to determine the personality traits of the individuals through classification. In an earlier work (Tutaysalgir, 2019), unsupervised learning methods have been used for the same problem. In this work, since supervised methods are more suitable, we decided to use informal Turkish text also with extended feature sets.

The main contribution of this paper can be described as designing a supervised learning framework for predicting personality traits of individuals from their informally written short Turkish texts, as tweets. The rich set of features obtained from the texts, their structures and the social network of the individuals are used in this framework. Some of these features are culture dependent and some others are language dependent. Thus, similar frameworks can be easily adapted for different languages.

In addition to conventional machine learning techniques, we used Long-Short Term Memory (LSTM) deep neural architecture, as well. In the literature, LSTM is shown to be very effective in sequential data processing. In many NLP applications statements are processed as a sequence of words and fed into LSTMs. In our problem we have developed LSTM based classifier to determine the personality trait scores of individuals by feeding the sentences of their tweets into LSTMs.

Furthermore, in order to verify our approach, we have also applied conventional classifiers to a very similar English texts, namely Facebook status updates, using a restricted form of our feature sets obtained from these texts. These experiments show that our approach produce similar results for both Turkish and English texts. Moreover, with our extended feature sets for Turkish data set that we have collected, we have obtained better results.

The contribution of the work can be summarized as follows:

- The number of studies on personality trait prediction on blog posts is limited. We further conduct the study on microblog posts in Turkish. Although the method is applicable to other languages, the results reported on Turkish texts have novelty since the applicability of the method on a morphologically complex language is revealed.
- In addition to the content features, network user statistics are included as features, as well.
- Various alternatives for feature extraction and vector construction have been studied in order to constitute the user profile.
- We further analyze the proposed method on informal English texts and reveal that the method including language independent feature set is applicable for English texts as well.
- Since the problem that we are dealing is a classification problem, we have investigated machine learning based classification methods in our work. For classification, in addition to applying conventional approaches including Support Vector Classifier (SVC), decision trees and nearest neighbor classifiers on the constructed vector model of the data, we modeled the problem as sequence labeling and applied Long-Short Term Memory (LSTM) deep neural architecture on the tweets of the users for personality prediction.

The rest of the paper is organized as follows: In Section 2, related studies in the literature are summarized. In Section 3, the details of the proposed method are presented. In Section 4, experiments are described and the results are discussed. The paper is concluded with an overview and future work in Section 5.

2. RELATED WORK

In the literature, different aspects of social network and different types of texts have been analyzed for personality trait prediction.

One of the aspects in social network to be considered for personality trait prediction is the number of friends of a user. This parameter is correlated with multiple traits of the five personality dimensions. In (Bachrach, Kosinski, Graepel, Kohli, & Stillwell, 2012), it is reported that the number of friends and neuroticism are negatively correlated. In (Schrammel, Köffel, & Tscheligi, 2009), the authors argue that more open people have more friends in social networks. The number of friends is also found to be correlated with extraversion in several studies (Amichai-Hamburger & Vinitzky, 2010), (Acar & Polonsky, 2007). In (Zalk et. al., 2010), it is reported that agreeable people are chosen as friends more often, since they are referred as easy to communicate. In social media, use of communication method is another aspect revealing the personality traits. In (Ross et. al., 2009) and (Moore & Mcelroy, 2012), it is shown that people who are more open to new experiences are more likely to post on others users' walls. In (Celli, 2011), it is reported that the users' tendency to retweet and The Big Five personality traits are correlated.

Linguistic signs obtained from social media usage are considered to be strong hints for recognizing personality (Mairesse, Walker, Mehl, & Moore, 2007). For instance, in (Golbeck, Robles, & Turner, 2011), linguistic features were half of the total features extracted from social media, and the highest number of correlations has been found for linguistic cues. Additionally, the authors included sentiment analysis results in their solution. It is reported that the most accurate predictions are obtained for Openness trait with ZeroR classifier. In (Tsytsarou & Palpanas, 2011), the correlation between sentiment scores of the texts and the personality traits of the authors is studied. In (Lima & De Castro, 2013), the authors include neural class in sentiment analysis, and propose a semi-supervised learning based solution. In (Quercia, Kosinski, Stillwell, & Crowcroft, 2011), 335 Twitter profiles were analyzed and these profiles are divided into 5 groups according to social media indices, such as listeners who tend to follow more users, popular people who have more followers. The authors worked on the correlation of personality dimensions and these five categories of microblog users.

In the literature, most of the studies on the use of linguistic cues focus on social media posts in English. One reason to this is the availability of data. Another reason is that most of the tools such Linguistic Inquiry and Word Count (LIWC¹) are built in English. However, there are efforts on other languages as well. Chinese is one of the most commonly studied languages about personality detection due to availability of data. In (Gao, ve diğerleri, 2013), social network posts in Chinese language are analyzed to discover personality of users. In order to extract features, LIWC is adapted for Chinese. The authors reported correlations between linguistic cues and music tastes of users. In another study on Chinese language (Wan, Zhang, Wu, & An, 2014), Naive Bayes and Logistic Regression are used for constructing personality prediction models. In (Peng, Liou, Chang, & Lee, 2015), feature selection is applied on a Chinese data set by using recursive feature elimination techniques. In (Arroju, Hassan, & Farnadi, 2015), tweets in various languages including English, Spanish, Dutch, and Italian are analyzed. Tokenized terms from the post content are matched with enhanced version of LIWC. It is reported that the best results achieved are in predicting the Openness dimension. Another study about Twitter texts is on Indonesian language (Pratama & Sarno, 2015). The authors extracted the most frequent 750 words similar to other open-vocabulary approaches. The classification algorithms used in the study are Naive Bayes, K-Nearest Neighbour (K-NN), and Support Vector Machine (SVM).

In a recent study (Sun et.al., 2019), Facebook data set, which we have also used to verify our approach, was used for personality prediction. The authors used classification based methods are used after applying word embeddings on the short texts in order to determine the personality traits of the individuals.

In another recent work, (Agrawal et.al., 2020), tweets in English has been used to predict the personality traits. The authors have also generated a framework similar to ours except that fewer

features have been used, and it is tested for English tweets. As supervised learning models, only traditional classifiers have been used.

Our work has similarity with the previous studies focusing on linguistic cues obtained from social media posts. However, in the literature, the analysis on texts in languages other than English is very limited and it has not been studied on Turkish texts before. Hence we address this gap in the literature, and one important difference of our work is that the analysis is conducted on posts in Turkish. Another important difference is the use of user statistics together with linguistic features. Furthermore, we model the problem as a sequence labeling problem and propose a LSTM based solution for personality prediction.

3. PROPOSED METHOD

The overall flow of the proposed method is presented in Figure 1. The process starts with data collection and ground truth construction for the analysis in Turkish (shown as light blue modules in the figure). The collected data is used within two learning models. In the first one (shown as green module), traditional supervised learning models such as Random Forest, SVC, linear SVC and K-NN classifier are used. Traditional classifiers can be used with any set of selected features of the problem. In our case one set is directly obtained from the tweets, and the other set requires word embedding process. In the first branch (shown as grey modules), the real effort and contribution lies in the vector construction step, which includes feature extraction from the tweets and using the words of the tweet content as the features (which are shown as two branches in the figure). For the second branch (shown as pink modules), firstly, important words are identified by using the Term Frequency - Inverse Document Frequency (TF-IDF) method, and then, they are mapped to points in high dimensional space using word embedding. We have made two sets of experiments with traditional classifiers. In the first one we have only used the features obtained from the first branch. In the second set of experiments, the outputs of these two separate phases for vector constructions are combined and used within traditional models. The last branch in the figure represents the second learning model through neural sequence labeling (shown in turquoise module). In this path, the tweets are given as input the neural architecture, LSTM. Note that, in this model, instead of textual representation, word embeddings of the words are constructed (shown as orange module). However, the word embedding construction is different from the one in the previous branch. After the model is trained by using the word embeddings, it maps the input tweet to a label of corresponding personality trait.

For traditional classifiers, each individual must be represented as a vector of selected and calculated features. Therefore, the process requires steps such as processing sentences written by individuals in order to collect many features. Also, important words are determined using TF-IDF and they are converted into vectors using word embedding techniques. However, in sequence label based model using LSTM the only process is to feed the sentences written by the users word by word into LSTM for the classification. Therefore, it does not require a detailed pre-processing phase and hence it is simpler to use.

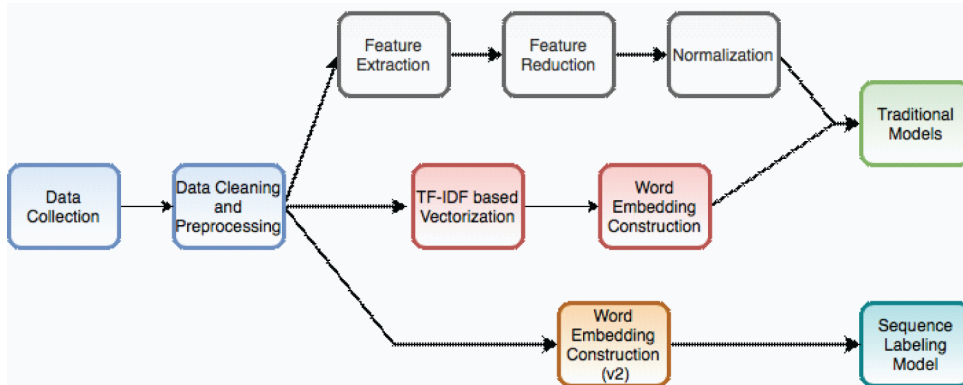
The details of the phases are described in detail in the following subsections.

3.1 Data Collection and Ground Truth Construction

Since we use supervised learning methods, in order to train our classifiers we needed ground truth for a set of inputs. In our problem, the input is the tweets of individuals and several additional information about the social network structure of the users. As a ground truth we need the actual values of personality traits of some individuals together with their profiles and their tweets.

Traditionally, personality is measured by using a set of multiple choice Likert type questions² (with choices as strongly agree, agree, neutral, disagree, and strongly disagree). The surveys that are used for this purpose generally contain 20 to 360 questions. The responses are used for calculating each trait's score separately for the user as a value between 0 to 100. Each choice contributes a predefined

Figure 1. Flow diagram of the proposed solution



value for each personality trait and the scores obtained from these surveys are considered as correct measures of the personalities of the participants (Costa & McCrae, 2012).

To have a valid ground truth, we needed accurate Big Five Personality traits results of the users whose tweets we are using. For this, we used an existing questionnaire, which automatically calculates accurate Big Five scores.

Our surveys are conducted as follows: Firstly, we translated the questionnaire to Turkish because the survey we have chosen was in English. The translation was done manually. Then, the survey is presented as an online form ³. The survey we used contains 60 questions, where each question has 5 possible choices. Gathering the entries of the user from the Turkish survey, the original survey is conducted in order to obtain the scores for each personality trait ⁴. The obtained result includes scores for each dimension of Big Five Personality Traits (Neuroticism, Openness, Extraversion, Agreeableness and Conscientiousnes). These values are represented as percentages. After we have obtained the scores of the individuals, we have discretized those scores into 4 groups as [0, 25), [25, 50), [50, 75), and [75, 100]. In the rest of the paper, we have used this discretization. In our data set, we did not have any instances for some score groupings of some personality traits. For example, among our volunteers there were no individual with “openness” score below 50. For the conducted survey, the score distribution for each dimension of Big Five Personality Traits is shown in Table 1.

Table 1. Big Five Score Dtribution

	0-25	25-50	50-75	75-100
<i>Openness</i>	0	0	15	25
<i>Conscientiousness</i>	0	13	18	9
<i>Extraversion</i>	4	18	12	6
<i>Agreeableness</i>	0	9	31	0
<i>Neuroticism</i>	0	10	23	7

In the survey, Twitter usernames of our volunteers are requested as well. By this way, we are able to construct the social network profiles of the volunteers and annotate them with Big Five personality traits’ scores. The most recent postings and the user statistics of the volunteers are gathered through

Twitter Search Application Programming Interface (API). It is important to note that, once the social network profile is constructed, we do not retain any identifying information, and hence the analysis is performed on anonymized data set. In general, we collected the following attributes of tweets:

- Textual content of tweet
- The date and time when the tweet is posted

In addition to tweets, we also collected the following statistical information for each user:

- Number of users followed
- Number of followers
- Number of tweets liked
- Number of tweets retweeted
- Number of total tweets
- Profile location
- Profile description
- Whether the user has default profile picture or not
- Whether the user has extended profile or not
- Whether the user has a profile background image or not

3.2 Pre-Processing

In the collected set of tweets, we applied a filtering to be able to retain only those that have higher potential to contribute to personality traits prediction. We aimed to obtain the texts written by the individuals as we assume that the textual features of the individuals contain information about their personalities. Therefore the tweets with the following characteristics are filtered out: Retweets of other tweets, tweets that are posted as an answer to another tweet, tweets that contain only URL, and tweets that mention about another user.

After this filtering, the amount of textual content is reduced, but the cleansed content has become more relevant for the analysis. In general, NLP tools are developed for processing formal texts, hence they are not exactly compatible with the messages posted in social media, which contain many linguistic irregularities and non-textual contents. Therefore, additional pre-processing is needed before analyzing such texts. To this aim, Twitter specific non-textual contents of URLs and mentions are removed. Additionally, multiple spaces, newlines and stop words are eliminated.

Additional pre-processings steps are also applied due to informal language use in tweets. The informal language can be due to both grammar and spelling mistakes, and use of slang words. For example, the tweet in Turkish *Bn kosmaya gidicem* (which is *I will go for a run* in English) has multiple spelling mistakes, which are frequent in informal language use. In order to handle such problems, we use SpellChecker functionality of Turkish morphological analysis tool Zemberek⁵. After spellchecking, the sample sentence is corrected as *Ben kořmaya gideceđim*.

It is also common to have Turkish words written by using English characters in tweets, and these words cause incorrect tagging as well. To correct these mistakes we used Zemberek's deasciification function. At the end of the extraction and the cleaning process, we have 8567 tweets.

3.3 Feature Extraction for Vector Construction

We used a set of features extracted from the use of language, timestamps of tweets, emoticons used within the tweets and several non-text properties, as described below, reflecting the user behavior:

- **Use of Language:** We collected 33 features from text including features on standard statistics such as word count. We count the words with different part of speech tags such as adjective,

verb and noun as well. There are also binary features such as sentiment orientation of the tweet. In order to calculate these features, we apply part of speech tagging by using Zemberek NLP tool. Then we calculate the average values of the tweets for each user to use as the weight of the feature. The word categories we used as feature are listed in Table 2. In this table each feature has a range corresponding to the average number of counts of that feature for individuals.

- **Timestamp:** Timestamp of tweets provides four features as Morning, Afternoon, Evening and Night. These features represent the frequency of the tweets which are posted in the specific time interval. Since these four time periods follow each other in a circular way, in order to calculate the distance between the instances more accurately, we represent these time periods by using two hot encoding. Two hot encoding is a special kind of bitwise representation for the feature values. With one hot encoding each pair of time intervals is treated as either equal or not. However, the evening is close to the night and the distance between the evening and the morning should be higher. In two-hot encoding, instead of having only a single 1 in a 4-element vector, where each element represents one time interval, two 1s are used such that one of them is representing that time interval and the other one is representing the previous (or the next) time interval. Thus, two consecutive time intervals will have one overlapping 1, and if they are not consecutive, then, there will be no overlapping 1s. In distance calculation, this approach will produce more accurate result when comparing time intervals.
- **Emoticons:** The search API that we used for collecting tweets of users outputs hexadecimal representations of emoticons, which is useful to understand the type of emoticon used. Firstly, we determined emoticons that can be useful for personality trait analysis. Then we created a dictionary with the most popular emoticons by grouping the related ones. We ended up with 11 groups of emoticons: Smiling, Affection, Tongue, Neutral, Unwell, Negative, Romantic, Fingers, Activity, Sport, and Plant. Each group in this dictionary represents a feature in our user vector. For each user, we count the number of emoticons used in each group and we calculate the average number of emoticon used by the user as a feature.
- **Non-Text Twitter User Information:** From Twitter, we obtain the following statistics for the users, and include them as features: Ratio of the number of users followed to the number of followers (followed/follower), the number of retweets, the number of total tweets.

3.4 Feature Elimination

To be able to retain only the most discriminative features, we used a distribution based analysis for features. We detected the features that are almost evenly distributed for all the users. To this aim, firstly, we calculated the variances for all the features. Then, the features with variance lower than the threshold value 0.01 are eliminated⁶. After feature elimination, we have the following 20 features: Evening, Night, Morning, Afternoon, Word, Adjective, Adverb, Noun, Verb, Plural, Full Stop, Pronoun, Incorrect, Punctuation, Numeral, Determiner, Conjunction, Negative, Negative Emoji, Smiling Emoji.

3.5 Normalization

Normalizing the feature values is an essential step in vector construction. We performed analysis with different normalization techniques, which are robust scaling, standard scaling and discretization. For robust scaling, and standard scaling, we used scikit-learn library⁷. For discretization, we used our own implementation of the algorithm.

Robust scaling is a robust method against outliers. For each feature, at the beginning, the first and the third quartiles and the median values are calculated. Then, each value is scaled as a ratio between its difference from the first quartile over the difference between the third and first quartile values. In standart scaling, firstly, the mean and the standard deviation values are calculated. Then,

Table 2. Features Extracted from the Use of Language

Feature	Range	Feature	Range
Case Ratio	0.58 – 1.0	Word	3.4 – 20.0
Verb	0.0 – 3.0	Noun	1.2 – 11.0
Punctuation	0.0 – 5.38	Adjective	0.0 – 2.5
Adverb	0.0 – 1.5	Numeral	0.0 – 2.06
Determiner	0.0 – 1.0	Post Positive	0.0 – 0.726
Duplicator	0.0 – 0.059	Conjunction	0.0 – 1.0
Interjection	0.0 – 1.0	Pronoun	0.0 – 1.23
Question	0.0 – 1.0	Incorrect	0.0 – 6.58
Negative	0.0 – 1.0	Plural	0.0 – 3.52
Present Time	0.0 – 1.0	Future Time	0.0 – 0.5
Past Time	0.0 – 0.88	Narrative Time	0.0 – 0.55
Progressive Time	0.0 – 1.5	Condition	0.0 – 0.5
Imperative	0.0 – 1.0	Necessity	0.0 – 0.42
Ability	0.0 – 0.5	Negative Ability	0.0 – 0.33
Question	0.0 – 1.0	Exclamation	0.0 – 0.67
Ellipsis	0.0 – 0.84	Full Stop	0.0 – 0.91
Non-Turkish Words	0.0 – 1.0		

z-score of each sample is calculated by subtracting the mean from it and then dividing the result by the standard deviation. In this method, outliers have more effect on the result than in robust scaling.

Discretization is one of the most frequently used forms of abstraction in machine learning. In some applications the actual value of a feature can be a continuous value from a predefined range, but, it can be processed more accurately if that range is split in a number of intervals. Thus, in the *discretization* method, the aim is to distribute the continuous space values into a selected number of bins. In our case, we discretize the values in 4 bins. To divide the data into sub-intervals, we use predefined threshold values. Then, we assigned corresponding size intervals as the new value.

In the validation analysis, among the three methods, discretization produced the highest accuracy results. Thus, as the normalization method, discretization is used in the rest of this study.

3.6 Term Frequency-Inverse Document Frequency (TF-IDF) Weighting

In addition to the basic features that are extracted from the collected tweets, we use the terms within the tweets of users as features as well. In order to determine the importance of the terms in the documents a standard statistical measure is used, called as Term Frequency-Inverse Document Frequency (TF-IDF). Term Frequency (TF) measures how frequently a term appears, and Inverse Document Frequency (IDF) measures how important it is. Basically, IDF is used to eliminate frequent but not important terms like “the”, “is” etc. Formally, they are defined as given Equations 1, 2 and 3:

$$TF(t) = \frac{\text{(Number of times term } t \text{ appears in a document)}}{\text{(Total number of terms in the document)}} \quad (1)$$

$$IDF(t) = \log e \frac{(Total\ number\ of\ documents)}{(Number\ of\ documents\ with\ term\ t\ in\ it)} \quad (2)$$

$$TF - IDF(t) = TF(t) * IDF(t) \quad (3)$$

Before applying TF-IDF weighting, the terms are pre-processed through tokenization and lemmatization.

Tokenization is the first step in NLP task, which extracts individual words from the sentences. Then, it is usually followed by lemmatization. The main aim in lemmatization is to replace the words with their lemmas in order to have the same word for the different forms of it. We use lemmas instead of stems because paragoges used in Turkish can change the word and stemming may not produce the proper form to represent the words. For example, *bardağım* (*my glass* in English) will turn into *bardağ* (not a correct form for glass) after stemming, but, lemmatization gives the result *bardak* (correct form of *glass*) and we use this correct form.

After the lemmatization, to be able to capture important words or phrases in the postings, we calculated the TF-IDF weights for 1-grams, 2-grams and 3-grams. Then, we retained only the phrases and words with highest TF-IDF weights. This process can also help us to learn topics the user tweeted about.

3.7 Word2vec Based Word Embedding

After determining the words with the highest TF-IDF values, we constructed word2vec embedding on them by using Gensim library⁸, which is a python library for topic modelling. The idea of word2vec embedding is to map words into high dimensional space such that the distance between the embeddings corresponds to the closeness of the words in natural language. Word2vec embeddings are generated using unsupervised learning methods on large corpuses. Our word2vec model has 38 dimensions, which is set through validation experiments. After validation experiments with different window sizes, we find that 7 is the optimum windows size in our case. Window size is the distance between guessed and present words in text. Additionally, we choose 3 as the minimum count. Hence, the words whose frequency is lower than 3 are ignored. After obtaining Word2vec representations of the top terms, we concatenate them with the features constructed in the previous steps. Hence for supervised learning model, we have vectors with 58 features per user.

3.8 Sequence Labeling Based Classification

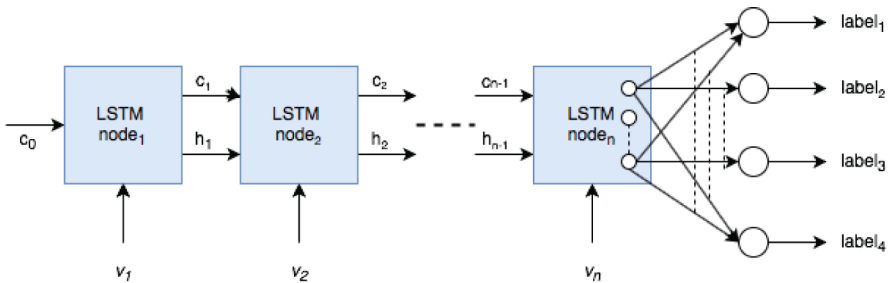
In addition to the vector based supervised learning for personality trait prediction, we also devised a sequence labeling based approach through a neural model directly on the textual content without feature extraction. In this method, as the input, each tweet of a user has been labeled by the user's Big Five Personality traits scores. In total, we have 17063 tweets from 40 users. Since the users' personality scores for all 5 dimensions have been obtained from the surveys we have annotated their tweets with these scores. The terms (tokens) in the tweets are represented with pre-trained word embedding vectors of 400 dimensions. If the pre-trained embedding is not available for a word, a random vector is selected uniformly in the range of [-0.25, 0.25].

To be able to encode the sequences, we used an LSTM neural model. LSTM was proposed as a new recurrent neural network architecture (Hochreiter, S. & Schmidhuber, J. 1997) in order to overcome the vanishing gradient problem of the conventional Recurrent Neural Networks (RNNs). In RNNs the errors between layers might vanish or blow up. This can causes convergence problems such as oscillating weights or very slow convergence. The first type of LSTM architecture had a mechanism

called as constant error carousel (CEC) in order to solve this problem. The memory cell of these LSTMs had input and output gates. The input gate is used for deciding which information should be kept or updated in memory cell, and the output gate is used to control which information should be output. Later, this standard LSTM was extended with a new feature called as forget gate, which is used for resetting the memory state that contains outdated information. In other words, this gate is responsible for resetting CEC. In addition to this, some more features, such as peephole connections and the full back-propagation through time (BPTT) were also added to LSTM architectures. We used this popular architecture, which is named as *Vanilla LSTM*.

To improve the robustness, we used drop-out technique with drop-out probability of 0.5. Hence, randomly chosen nodes have been removed automatically. We have observed that the employed drop-out rate prevented overfitting. For classification, output of LSTM is fed into a fully-connected layer with four output nodes (which is the number of the labels). As the activation function, the softmax function has been used. Softmax function is a frequently used activation function for classification tasks in deep neural architectures. It generates a vector of probability distributions from the numerical values of multi-class classification ($label_1, label_2, label_3$ and $label_4$, which correspond to 4 score grouping of personality traits, in our case). Usually special logarithmic functions are used for the final activation function in order to handle any range of numerical values. The general LSTM neural architecture used is shown in Figure 2. The main input of LSTM architecture is the sequence of word embeddings corresponding to the words of the tweets (as v_1, v_2, \dots, v_n in the figure). Each cell also gets c_i and h_i as an input from the previous cell. The first cell gets the input vector c_0 , which could be only 0, and it is transformed to the next cell with small modifications, since it carries the state information. Each cell also produces an output h_i , which are fed as input to the next cell. The final output, on the other hand, is converted into the labels of the classifier.

Figure 2. LSTM Architecture



4. EXPERIMENTS

We report the accuracy results in two ways. In the first one, the accuracy is obtained on the training samples. In the second one, 3-fold cross validation results are reported. Our data set has 40 individuals with personality scores obtained from the surveys they have taken. We had collected more than 17000 tweets of these individuals.

4.1 Classification Performance Analysis With TF-IDF and Word2Vec Features Only

To analyze the effect of TF-IDF weighting and word2vec embedding features for personality detection accuracy, we firstly applied classification involving only these features. In this experiment, in addition to our Turkish dataset, we also used another dataset in English⁹. This dataset has been used in several studies and promising results have been obtained (Youyou, Kosinski, & Stillwell, 2015). The dataset

includes Facebook status updates of 251 users. Additionally, it contains Big Five personality scores of the users. As in the word embedding obtained for the tweets in Turkish data set, we extracted the textual content of tweets in English and obtained word2vec vectors. This dataset is very similar in terms of the structure to Turkish Twitter dataset, and both are annotated similarly with personality traits scores. However, several features that we have mentioned earlier related to the Turkish Tweet data set did not exist in this data set. Therefore, we had to use only a limited version of those features. With these experiments, we wanted to show that when similar features are used on two different data sets we can obtain similar results.

Hence, the steps followed to process the Facebook English dataset is similar to those applied for the Twitter Turkish dataset. Firstly, we cleaned the data from double spaces and punctuation marks. Then, by using Porter Stemmer ¹⁰, tokens are lemmatized. By using scikit-learn's TF-IDF counter, we find the mostly used words for each user. To fulfill this task, stop words are eliminated, and not only single words but also 2-grams and 3-grams are tokenized as well. After sorting the tokenizations according to their TF-IDF scores, we determine the top 20 words for each user. To create a feature vector from the top 20 words, we used a pre-trained word embedding corpus for English, which has been created from Wikipedia ¹¹. In this collection, each word is represented by a vector of 300 features.

For classification, we used 4 different algorithms, namely Random Forest Classifier, SVC, Linear SVC, K-NN classifier.

The comparison of personality traits prediction accuracy of Turkish and English word2vec features is shown in Table 3 in terms of accuracy on the training data and 3-fold cross validation. We have presented these results also just to show that the models trained for two different data sets produce similar outcome. Note that we analyze and compare the accuracy results in terms of 3-fold cross validation. We can summarize our observations on the evaluation results as follows:

- The results obtained on Turkish social media content have similar accuracy to that of English data set. This gives the clue that the proposed approach is applicable to data sets in languages having different structures.
- As the supervised learning model, SVC provides better results compared to the other classification algorithms. This observation is valid for both English and Turkish social media data sets.
- The prediction performance on Turkish and English data set presents variation according to the personality traits. For Neuroticism, prediction accuracy is considerably better for English social media data set, whereas for Turkish data set, for the other personality traits, the prediction performance is higher than that on English data set.

The differences between two datasets is most probably due to sizes of these datasets, since Turkish data set is around the size of 20% of the English data set. The variations in personality traits observation distributions may be another reason for differences in the prediction accuracies for different personality traits. However, the results also show that our dataset can be used for evaluating the personalities from Turkish texts as well with acceptable error rates.

4.2 Classification Performance Analysis Under All Features

We perform classification on Turkish data in another configuration as well. We have added the features coming from Twitter analysis to the TF-IDF and word2vec features. These evaluation results are shown in Table 4.

Similar to the previous experiments, for this one, we also present the training accuracy results and 3-fold cross validation accuracies. When the results of extended feature set version of Turkish dataset is compared to the one having only word2vec features, the observations can be summarized as follows:

Table 3. Prediction Accuracy by Word2Vec Features

Openness	Training Accuracy (English)	Training Accuracy (Turkish)	3-Fold Cross Validation (English)	3-Fold Cross Validation (Turkish)
<i>Random Forest Classifier</i>	0.98	0.93	0.43 (+/- 0.09)	0.56 (+/- 0.24)
<i>SVC</i>	0.49	0.67	0.5 (+/- 0.02)	0.67 (+/- 0.05)
<i>LinearSVC</i>	1.0	0.66	0.42 (+/- 0.07)	0.67 (+/- 0.05)
<i>K-NN Classifier</i>	0.50	0.59	0.36 (+/- 0.19)	0.46 (+/- 0.28)
Conscientiousness	Training Accuracy (English)	Training Accuracy (Turkish)	3-Fold Cross Validation (English)	3-Fold Cross Validation (Turkish)
<i>Random Forest Classifier</i>	0.97	0.93	0.47 (+/- 0.17)	0.52 (+/- 0.29)
<i>SVC</i>	0.61	0.63	0.44 (+/- 0.07)	0.63 (+/- 0.08)
<i>LinearSVC</i>	1.0	0.63	0.42 (+/- 0.09)	0.63 (+/- 0.08)
<i>K-NN Classifier</i>	0.57	0.46	0.37 (+/- 0.09)	0.44 (+/- 0.17)
Extraversion	Training Accuracy (English)	Training Accuracy (Turkish)	3-Fold Cross Validation (English)	3-Fold Cross Validation (Turkish)
<i>Random Forest Classifier</i>	0.99	0.96	0.47 (+/- 0.05)	0.53 (+/- 0.3)
<i>SVC</i>	0.66	0.53	0.51 (+/- 0.1)	0.53 (+/- 0.05)
<i>LinearSVC</i>	1.0	0.8	0.48 (+/- 0.04)	0.8 (+/- 0.12)
<i>K-NN Classifier</i>	0.66	0.43	0.44 (+/- 0.13)	0.46 (+/- 0.18)
Agreeableness	Training Accuracy (English)	Training Accuracy (Turkish)	3-Fold Cross Validation (English)	3-Fold Cross Validation (Turkish)
<i>Random Forest Classifier</i>	0.99	0.96	0.46 (+/- 0.07)	0.5 (+/- 0.43)
<i>SVC</i>	0.55	0.8	0.53 (+/- 0.05)	0.8 (+/- 0.12)
<i>LinearSVC</i>	1.0	0.56	0.48 (+/- 0.04)	0.8 (+/- 0.12)
<i>K-NN Classifier</i>	0.5	0.63	0.37 (+/- 0.08)	0.57 (+/- 0.26)
Neuroticism	Training Accuracy (English)	Training Accuracy (Turkish)	3-Fold Cross Validation (English)	3-Fold Cross Validation (Turkish)
<i>Random Forest Classifier</i>	0.99	0.90	0.69 (+/- 0.07)	0.42 (+/- 0.24)
<i>SVC</i>	0.74	0.53	0.75 (+/- 0.02)	0.54 (+/- 0.12)
<i>LinearSVC</i>	1.0	0.56	0.62 (+/- 0.11)	0.54 (+/- 0.12)
<i>K-NN Classifier</i>	0.64	0.36	0.48 (+/- 0.2)	0.35 (+/- 0.24)

Table 4. Prediction Accuracy for Turkish Tweets

Openness	Training Accuracy	3-Fold Cross Validation
<i>Random Forest Classifier</i>	1.0	0.54 (+/- 0.18)
<i>SVC</i>	0.66	0.67 (+/- 0.05)
<i>LinearSVC</i>	0.66	0.47 (+/- 0.05)
<i>K-NN Classifier</i>	0.4	0.44 (+/- 0.11)
Conscientiousness	Training Accuracy	3-Fold Cross Validation
<i>Random Forest Classifier</i>	0.96	0.46 (+/- 0.28)
<i>SVC</i>	0.53	0.53 (+/- 0.05)
<i>LinearSVC</i>	0.66	0.50 (+/- 0.14)
<i>K-NN Classifier</i>	0.5	0.42 (+/- 0.34)
Extraversion	Training Accuracy	3-Fold Cross Validation
<i>Random Forest Classifier</i>	0.96	0.32 (+/- 0.31)
<i>SVC</i>	0.53	0.54 (+/- 0.12)
<i>LinearSVC</i>	0.76	0.62 (+/- 0.39)
<i>K-NN Classifier</i>	0.43	0.28 (+/- 0.33)
Agreeableness	Training Accuracy	3-Fold Cross Validation
<i>Random Forest Classifier</i>	0.96	0.83 (+/- 0.09)
<i>SVC</i>	0.80	0.80 (+/- 0.07)
<i>LinearSVC</i>	0.90	0.83 (+/- 0.09)
<i>K-NN Classifier</i>	0.80	0.77 (+/- 0.09)
Neuroticism	Training Accuracy	3-Fold Cross Validation
<i>Random Forest Classifier</i>	0.96	0.57 (+/- 0.05)
<i>SVC</i>	0.63	0.63 (+/- 0.05)
<i>LinearSVC</i>	0.63	0.57 (+/- 0.24)
<i>K-NN Classifier</i>	0.30	0.17 (+/- 0.10)

- When compared under 3-fold cross validation results, the performances vary with respect to personality traits. For Agreeableness and Neuroticism, additional features have positive effect on the accuracy, whereas for the others, we observe a slight drop. This may be due to that extracted features, such as the number of retweets and the number of words, carry stronger cues for Agreeableness and Neuroticism personality traits.
- In this experiment, SVC and Linear SVC provide the highest accuracy results as in the previous experiment.

4.3 Classification Performance Analysis With Sequence Labeling

Sequence labeling is based on LSTM, and its analysis is conducted on tweets of the users under 3-fold cross validation. The partitioning is applied on the basis of the users. In other words, the user collection is partitioned into three, and each partition consists of the tweets of the corresponding users in it. The accuracy is calculated on the basis of the correctness of tweet label's prediction. The result of the analysis is given in Table 5.

Table 5. Prediction Accuracy with Sequence Labeling for Turkish Tweets

Personality Traits	Accuracy under 3-Fold Cross Validation
Openness	0.56 (+/- 0.12)
Conscientiousness	0.47 (+/- 0.04)
Extraversion	0.50 (+/- 0.18)
Agreeableness	0.88 (+/- 0.04)
Neuroticism	0.69 (+/- 0.07)

As seen in the results, the accuracy performance is similar to those given in Table 4. For example, the highest accuracy result is obtained for the “agreeableness” from both the traditional classifiers and LSTM classifier. Similarly, both models produce the lowest accuracy on the “conscientiousness”. Sequence labeling with LSTM neural model gives slightly better results for Neuroticism and Agreeableness, whereas the accuracy is lower for Openness, Conscientiousness and Extraversion. However, the general behavior with respect to personality traits is similar.

5. CONCLUSION

In this work, we present a method for personality traits prediction by using Twitter data. We have designed our framework to work on Turkish tweets, however, the whole process can be applied on any language by excluding language specific features and using language specific tools such as morphological analyzers.

The proposed method is composed of the steps of personality traits annotation through Big Five survey, social media profile construction and prediction model construction. The novelty of the work basically lies in two steps of the process. The first one is the social media profile construction, in which we elaborated on the use of linguistic features extracted from posts, features from social media user statistics such as the number of tweets, and retweets. Additionally, we analyzed effect of using word embeddings for linguistic features. As the second contribution, we consider the textual content as the essence of the profile, consider the problem as sequence tagging and apply a deep neural model as the classifier.

In the experiments, we analyze the effect of textual features on two data sets, one in Turkish and the other in English. Although the size of the Turkish data set is limited, we have observed similar prediction performances on both Turkish and English data sets. This gives a strong clue for the applicability of the approach in morphologically complex languages as well.

For the analysis on the effect of additional features extracted from the tweets, it is observed that they have positive effect especially for Agreeableness and Neuroticism personality traits. This indicates that, in addition to the word use in the language, the network and communication patterns are strong cues for some of the personality traits. As for prediction performance of the neural model on tweets, we observe a similar performance to the vector based classification. The similarity is in both the prediction accuracy per personality trait dimension, and the general prediction performance. The results indicate that the neural approach has potential to be used for personality prediction.

As a future work, it is possible to extend the feature set by including more features from the network structure of individuals whose tweets are analysed. In this work, we have only used some features that can be obtained from user profile directly. However, there is a very rich additional information that can be obtained from the network structure, such as the follower and followee links of the users, also called as ego network. The structures of the nodes in this network can also provide new features that can be useful in analysing the personality of individuals.

REFERENCES

- Acar, A. S., & Polonsky, M. (2007). Online Social Networks and Insights into Marketing Communications. *Journal of Internet Commerce*, 6(4), 55–72. doi:10.1080/15332860802086227
- Ahmad, H., Asghar, M., Khan, A., & Habib, A. (2020). A Systematic Literature Review of Personality Trait Classification from Textual Content. *Open Computer Science*, 10(1), 175–193. doi:10.1515/comp-2020-0188
- Amichai-Hamburger, Y., & Vinitzky, G. (2010). Social Network Use and Personality. *Computers in Human Behavior*, 26(11), 1289–1295. doi:10.1016/j.chb.2010.03.018
- Arroju, M., Hassan, A., & Farnadi, G. (2015). Age, gender and personality recognition using tweets in a multilingual setting. *CLEF 2015 working notes*.
- Ashton, M., Lee, K., & Paunonen, S. V. (2002). What Is the Central Feature of Extraversion? Social Attention Versus Reward Sensitivity. *Journal of Personality and Social Psychology*, 83(08), 245–252. doi:10.1037/0022-3514.83.1.245 PMID:12088129
- Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012). Personality and Patterns of Facebook Usage. *Proceedings of the 3rd Annual ACM Web Science Conference*. doi:10.1145/2380718.2380722
- Baddeley, J., & Singer, J. (2009). A loss in the family: Silence, memory, and narrative identity after bereavement. *Memory (Hove, England)*, 18(09), 198–207. PMID:19697249
- Bharadwaj, S., Sridhar, S., Choudhary, R., & Srinath, R. (2018). Persona Traits Identification based on Myers-Briggs Type Indicator (MBTI)- A Text Classification Approach. *Proceeding of International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1076-1082. doi:10.1109/ICACCI.2018.8554828
- Boele, R. (2000). *The Big Five Personality Factors: The psycholexical approach to personality*. Hogrefe & Huber Publishers.
- Celli, F. (2011). *Mining User Personality in Twitter*. Language, Interaction and Computation Laboratory (CLIC).
- Costa, P., & McCrae, R. R. (2012). The Five-Factor Model, Five-Factor Theory, and Interpersonal Psychology. In *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions* (pp. 91-104). John Wiley and Sons.
- Fast, L., & Funder, D. (2008). Personality as Manifest in Word Use: Correlations With Self-Report, Acquaintance Report, and Behavior. *Journal of Personality and Social Psychology*, 94(03), 334–346. doi:10.1037/0022-3514.94.2.334 PMID:18211181
- Gao, R., Hao, B., Bai, S. L. L., Li, A., & Zhu, T. (2013). Improving user profile with personality traits predicted from social media content. In *Proceedings of the 7th ACM Conference on Recommender Systems* (s. 355–358). ACM. doi:10.1145/2507157.2507219
- Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *Proceedings of the Conference on Human Factors in Computing Systems*, 253–262.
- Goldberg, L. R. (1990). An alternative “description of personality”: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216–1229. doi:10.1037/0022-3514.59.6.1216 PMID:2283588
- Graziano, W., & Eisenberg, N. (1997). Agreeableness: A Dimension of Personality. In *Handbook of Personality Psychology* (pp. 795–824). Academic Press.
- Hirsh, J., & Peterson, J. (2009). Extraversion, neuroticism, and the prisoner’s dilemma. *Personality and Individual Differences*, 46(12), 254–256. doi:10.1016/j.paid.2008.10.006
- Lepine, J. A., Colquitt, J., & Erez, A. (2006). Adaptability to changing task contexts: Effects of general cognitive ability conscientiousness, and openness to experience. *Personnel Psychology*, 53(12), 563–593.
- Lima, A. C., & De Castro, L. (2013). Multi-Label Semi-Supervised Classification Applied to Personality Prediction in Tweets. *Proceedings of the First BRICS Countries Congress on Computational Intelligence*. doi:10.1109/BRICS-CCI-CBIC.2013.41

- Mairesse, F., Walker, M., Mehl, M., & Moore, R. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30(09), 457–500. doi:10.1613/jair.2349
- McCrae, R. R., & Costa, P. (1997). Personality Trait Structure as a Human Universal. *The American Psychologist*, 52(6), 509–516. doi:10.1037/0003-066X.52.5.509 PMID:9145021
- Mehl, M., Gosling, S., & Pennebaker, J. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(06), 862–877. doi:10.1037/0022-3514.90.5.862 PMID:16737378
- Moore, K., & Mcelroy, J. (2012). The Influence of Personality on Facebook Usage, Wall Postings, and Regret. *Computers in Human Behavior*, 28(01), 267–274. doi:10.1016/j.chb.2011.09.009
- Odekerken, G., De Wulf, K., & Schumacher, P. (2003). Strengthening outcomes of retailer–consumer relationships: The dual impact of relationship marketing tactics and consumer personality. *Journal of Business Research*, 02(3), 177–190. doi:10.1016/S0148-2963(01)00219-3
- Peng, K., Liou, L., Chang, C., & Lee, D. (2015). Predicting personality traits of Chinese users based on Facebook wall posts. *24th Wireless and Optical Communication Conference (WOCC)*, 9-14. doi:10.1109/WOCC.2015.7346106
- Pennebaker, J., & King, L. (2000). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(1), 1296–1312. PMID:10626371
- Pratama, B. Y., & Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. *International Conference on Data and Software Engineering (ICoDSE)*, 170-174. doi:10.1109/ICODSE.2015.7436992
- Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 180-185. doi:10.1109/PASSAT/SocialCom.2011.26
- Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2), 578–586. doi:10.1016/j.chb.2008.12.024
- Rozin, P. (2001). Social Psychology and Science: Some Lessons From Solomon Asch. *Personality and Social Psychology Review*, 5(2), 2–14. doi:10.1207/S15327957PSPR0501_1
- Schrammel, J., Köffel, C., & Tscheligi, M. (2009). Personality Traits, Usage Patterns and Information Disclosure in Online Communities. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (pp. 167-174). British Computer Society. doi:10.14236/ewic/HCI2009.19
- Suhartono, D., Ong, V. R., S., A. D., Aryo, W., Nugroho, . . . Suprayogi, M. N. (2017). Personality Prediction Based on Twitter Information in Bahasa Indonesia. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, 11, 367–372. doi:10.15439/2017F359
- Tsytasarau, M., & Palpanas, T. (2011). Mining Subjective Data on the Web. *Data Mining and Knowledge Discovery*, 24(05), 478–514.
- Wan, D., Zhang, C., Wu, M., & An, Z. (2014). Personality Prediction Based on All Characters of User Social Media Information. *Communications in Computer and Information Science*, 489, 220–230. doi:10.1007/978-3-662-45558-6_20
- Widiger, T., & Cost, P. T. (2012). *Personality disorders and the Five - Factor Model of Personality*. American Psychological Association.
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, 112(4), 1036–1040. doi:10.1073/pnas.1418680112 PMID:25583507
- Zalk, M., Burk, W., Branje, S., Denissen, J., Aken, M., & Meeus, W. (2010). Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach. *Journal of Personality*, 78(04), 509–538. PMID:20433629
- Zhang, L. (2002). Thinking Styles and the Big Five Personality Traits. *Educational Psychology*, 22(01), 17–31. doi:10.1080/01443410120101224

ENDNOTES

- ¹ <http://liwc.wpengine.com>
- ² Likert scale is the most widely used approach to scaling responses in survey research. The scale items denote to the level the user agrees with the expression given in the question, such as *strongly agree*, *agree*, *neutral*, *disagree* and *strongly disagree*.
- ³ We used Google Forms as the online form (<https://www.google.com/intl/en-GB/forms/about/>)
- ⁴ The original survey is run as a Selenium application by using the collected entries (<https://www.seleniumhq.org>)
- ⁵ <https://github.com/ahmetaa/zemberek-nlp>
- ⁶ Threshold value is determined through validation experiments. For this task, we used Scikit-learn's VarianceThreshold function (https://scikit-learn.org/stable/modules/feature_selection.html).
- ⁷ <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>
- ⁸ <https://pypi.org/project/gensim/>
- ⁹ <https://sites.google.com/michalkosinski.com/mypersonality>
- ¹⁰ <https://github.com/jedijulia/porter-stemmer>
- ¹¹ <https://www.kaggle.com/yesbutwhatdoesitmean/wikinews300d1mvec>

Gökalp Maviş is a computer engineer in MSc degree, currently working in Siemens. His passion in computer engineering is machine learning and data science. His thesis is about personality assessment by using Twitter data.

Ismail Hakki Toroslu (PhD) is with the Department of Computer Engineering, Middle East Technical University (METU) since 1993. He has received his B.S. and M.S. degrees in computer engineering from METU, Ankara in 1987 and Bilkent University, Ankara in 1989 respectively. Prof. Toroslu received his PhD from the Department of Electrical Engineering and Computer Science at Northwestern University, IL, in 1993. He has been a visiting professor in the Department of Computer Science at University of Central Florida between 2000 and 2002. His current research interests include data mining, sentiment analysis and text mining, and recommendation systems. Prof. Toroslu has published more than 80 technical papers in variety of areas of computer science. Prof. Toroslu has also received IBM Faculty Award in 2010.

Pinar Karagoz (PhD) is currently Professor in Computer Engineering Department, Middle East Technical University (METU). She received her Ph.D. from the same department in 2003. She worked as a visiting researcher in State University of New York (SUNY) at Stony Brook. Her research interests include data mining, web usage mining, social network analysis, information extraction from the web, semantic web services, web service discovery and composition. Dr. Karagoz has authored several publications in international journals and leading conferences. Some of her papers were published in journals such as IEEE TKDE, IEEE Industrial Informatics, ACM TWEB, Information Systems Journal, SIGMOD Record, Knowledge and Information Systems, Knowledge based Systems. Some of her research were presented and published in conferences including VLDB, CIKM, ASONAM, DAWAK, ICWS. She has also taken part in the organization committee of several conferences including ICDM and VLDB.