Twitter Verisi ve Makine Öğrenmesi Modelleriyle Kişilik Tahminleme

Predicting Personality with Twitter Data and Machine Learning Models

İzel Ergu
Computer Engineering Department
Dokuz Eylül University
İzmir, Turkey
izel.ergu@ceng.deu.edu.tr

Zerrin Işık
Computer Engineering Department
Dokuz Eylül University
İzmir, Turkey
zerrin@cs.deu.edu.tr

İsmail Yankayış
Computer Engineering Department
Dokuz Eylül University
İzmir,Turkey
ismail.yankayis@cs.deu.edu.tr

Özet— **Bu çalışma, kullanıcı tarafından gönderilen tweet'lerde kullanılan kelimelerle Twitter kullanıcısının kişilik tahminini yapmaktadır. Kullanıcıların kişiliği, uyumluluk, sorumluluk, açıklık, nevrotiklik ve dışa dönüklük kişilik özelliklerini tanımlayan Big Five Personality Model'e dayanarak tahmin edilir. Tahminleme için Türkçe kelimeleri analiz ettik ve özel kelime grupları ile Türkçe kelimeleri içeren yeni bir sözlük hazırladık. Her bir kişilik özelliğini tahmin etmek için en başarılı makine öğrenmesi yöntemleri seçilmiştir. Makine öğrenmesi modelleri kullanıcı tarafından atılan son 50 tweet ile eğitildiğinde, her bir kişilik özelliğini 0.76 ila 0.97 aralığında doğruluk değerleri ile tahmin etmektedir.**

Anahtar kelimeler— sosyal medya, Twitter, kişilik, makine öğrenmesi, Türkçe metin

Abstract— **This study applies personality prediction of a Twitter user based on the words used in tweets posted by the user. The personality type is predicted based on Big Five Personality Model that outputs agreeableness, conscientiousness, openness, neuroticism, and extraversion as personality traits. We analyzed Turkish words for prediction, prepared a new dictionary that includes Turkish words with their special word groups. The most successful machine learning methods are selected to predict each personality trait. When the machine learning models were trained with the latest 50 tweets of users, models estimated each personality trait with the accuracy values in the range of 0.76 to 0.97.**

Keywords— social media, Twitter, personality, machine learning, Turkish text

I. INTRODUCTION

Nowadays, social media is the most popular environment among people. Most of the people use social media to share their emotions, daily life activities, ideas about several events (e.g., political, agenda topics) in the form of photos or texts. Twitter is one of the most widely used microblogging and social networking services that reflects people's emotions as texts. If somebody is angry, happy or sad about an event, he or she generally prefers sharing these feelings by posting a tweet. It is a tool for people to show a reaction to events. It does not only include texts, but there are also images that reflect the situation. This study only analyzed the text content of tweets. People usually do not think about hiding their personality. It means that they do not act like somebody else. Therefore, it provides us an opportunity for predicting the personality of a person from his/her shared texts.

This study aims to predict the personality of a user by using an intelligent personality assessment model. A personality model that is called Big Five Personality Model is used for the assessment of the proposed model. In this study, Turkish tweets are collected and preprocessed before applying machine learning models. A new dataset was prepared with the test results and tweets of 51 volunteer Twitter users. Four different machine learning models, which are the most successful ones, are used for classifying five personality traits. To the best of our knowledge, this study is the first research to predict personality traits by using Turkish tweets and machine learning models.

II. RELATED WORKS

There are many researches about personality prediction. Qui et al. applied a personality prediction from microblogs of users [1]. This study had three folds: (1) personality prediction based on microblogs, (2) detection of linguistic cues related with personality traits, (3) identification of potential linguistic cues that is identified by observers for personality prediction. A total of 28,978 tweets were collected for study. Big five

978-1-7281-2868-9/19/$31.00 ©2019 IEEE

personality traits method was used for prediction. The tool called Linguistic Inquiry and Word Count (LIWC) was used to identify linguistic patterns related with personality traits [2]. Linguistic cues were found for each personality traits. After linguistic cues were determined, eight human experts processed tweets without any time restriction and rated their opinions about the participants' personality. As a result, only conscientiousness personality trait was found not to correlate significantly with any cues and found that cues are negatively or positively related with other personality traits.

Another study about personality prediction was done in University of Maryland [3]. Firstly, a test with 45-questions was administered to users, then at least 2,000 tweets of each user were collected. Some statistics were collected for each user with respect to tweets and questionnaire. These are number of followers, number of followings, density of the social network, number of mentions, number of replies, number of hashtags, number of links and words per tweet. The primarily analysis was the processing of text in tweets, then linguistic analysis was applied to tweets. Again, the LIWC tool was used to analyze the content of tweets. In addition, word by word sentiment analysis was performed on 8 tweets of each user. WEKA was used for regression analysis. Gaussian Process and ZeroR were used with 10-fold cross validation. Two of them gave similar results. The Mean Absolute Error on a normalized scale were calculated for each personality trait. The results of ZeroR for agreeableness, conscientiousness, extraversion, neuroticism and openness are 0.129, 0.146, 0.160, 0.182, and 0.119, respectively. The results of Gaussian process for agreeableness, conscientiousness, extraversion, neuroticism and openness are 0.130, 0.145, 0.160, 0.182, and 0.119, respectively.

Word embedding with Gaussian Process was proposed in another study [4]. Their method is efficient in short text classification. To collect data, a survey was administered to users and 10 to 200 tweets of users were fetched. Words from the tweets were extracted and their word embeddings are represented with a vector. Gaussian processes model took this vector as an input for training. 3-gram and LIWC with Ridge Regression (RR) were used to compare methods. As a result, the new method is 33% better than the previous best method and 3-gram RR features gave better results than LIWC features.

III. METHODS

A. *Tools*

The Twitter API platform provides wide access to public Twitter data in which users share their tweets. It allows to access tweets according to the username with the limit of 200 tweets. The "Tweepy" library in python is a widely used to access Twitter data. It has its own functions to access all kinds of information of public accounts. In this project, Zemberek-NLP is used for word operations in Turkish and it is a Java library.

LIWC is a naive program that reads given text and computes the frequency of the words that occur in the given text. It includes fundamental text analysis modules. LIWC has two features: the processing component and the dictionary. The processing component takes different type of files as an input, and then computes the word frequencies of each file. The dictionary refers to the collection of words. It is compared with each file to identify associated psychological category of each word. LIWC counts every word in the given text and returns the number of each category. It reports all categories and their amount as the output of the processing [2].

LIWC is commonly used in psychological domain, since it counts words based on psychological meaning word categories. Although, it counts words in psychological categories, the result of experiments using LIWC has the ability to detect meaning in a wide variety such as attentional focus, emotionality, social relationship, thinking styles, and individual differences.

B. *Data Collection*

Twitter API is used to collect the efficient number of tweets in this project. While collecting tweets, Turkish language for tweets and username filters are applied.

We need personality cues to predict the level of suitability for each personality traits. To calculate correlations between each personality cues and each personality traits, we need to have actual test results that are answered by different real Twitter users. There were 51 volunteers who are active Twitter users and answered the Big Five Personality test. If test the score of a person is lower than or equal to 15 for a specific trait, this person is not suitable at all for this personality trait; if the score is between 16 and 20, this person is not suitable for this personality trait; if the score is between 21 and 30, this person is suitable for this personality trait; if the score is higher than or equal 31, this person is very suitable for this personality trait.

There is a corpus in a previous study [4] that covers the most commonly used Turkish words used in Twitter. It contains 103,000 words in total. We used this corpus to rearrange the LIWC dictionary. Since a new dictionary with as many Turkish words as possible should be included to prepare the most suitable dictionary and to obtain the most successful result in the project. After collecting words from this corpus, they should be preprocessed to use.

C. *Data Preprocessing*

Collected tweets are in the JSON format. It includes various kinds of information that we might not need during our analysis. We only need the word content in tweets, but they include retweets or links. Besides these, people prefer to use daily language, abbreviations for some words or they misspell some words unintentionally. Before applying word stemming, such irregularities should be resolved. Before using morphological analyzer, some items must be removed from the text of tweet. If text contains links that must be removed, because this study aimed that personality was assessed by using words not using videos, pictures or contents. After all unnecessary items are removed, the next step is stemming by using the Zemberek library.

In this study, the Zemberek library is used to find stems and suffixes of Turkish words. There is a class called morphology that provides morphological analysis, morphological ambiguity resolution, and word generation; it also contains its own normalizer and analyzer methods. Normalizer method provides correcting misspelled words and analyzer method provides stemming.

The LIWC dictionary is developed for English words, however it is not available for Turkish words. We used the “googletrans” library in python to translate each English word in this dictionary to Turkish. There are 4,328 words that are translated from English to Turkish. After their translation, each word is controlled for correctness of translation and the suitability between the meaning of each word and the groups. There are some words that need to be added or deleted, also some words that need to be added to default dictionary of morphology class.

Commonly used Turkish words were processed by using the Zemberek stemming method. Before stemming, there were more than one word with the same root that has the inflection. We removed these words and then stemming was performed. As a result of stemming, there are 10.000 words in our corpus. We used this corpus to complete the missing words in the Turkish LIWC dictionary.

The initial English dictionary includes 4,568 words. After we translated them into Turkish, we obtained few more words than in the original dictionary. To complete it, we add commonly used Turkish words to this dictionary. While we were adding commonly used Turkish words to the dictionary, we looked original grouping of LIWC dictionary as a reference. On the other hand, there are many ambiguities in Turkish words. It means that one word can represent more than one meaning. So, in such situations, we added categories of both meanings to the dictionary for these words. Finally, we have a dictionary with 5276 words that is used to analyze words.

According to related works, the LIWC library is used to find out to group of words in similar studies. The “liwc” library in python was used the LIWC dictionary. We used this library to find out the group of a given word. After stems are inserted to database, the “liwc” script returns stem from database and displays word groups with their counts according to the dictionary from given tweets. We will use these word groups and counts as the input vectors of machine learning algorithm.

D. Dataset Preparation

Dataset contains frequencies of each word groups and personality test scores of each user. Each tuple has frequencies of 41 different word groups and the scores of 5 different personality traits. There are 51 instances (persons) in our final dataset.

The first step of data preparation is normalizing the observed frequencies of word groups. The normalization was done by Equation 1 in which the actual value is divided by total number of grouped words. Before applying normalization, the frequency values are very different from each other and the dataset has uneven distribution. To obtain a normal distribution, this operation was applied. Normalization was applied as row (tuple)-based.

$$\mathrm{a}_i' = \frac{\mathrm{a_i}}{\Sigma \mathrm{a_i}} \qquad (1)$$

We reduced the number of word groups by applying a correlation analysis. Before finding correlation between them, we performed a range normalization for each word group by using Equation 2. After applying a range normalization, the frequencies of word groups lie between [0,1].

$$\mathrm{a}_i' = \frac{\mathrm{a_i} - \min(\mathrm{a})}{\max(\mathrm{a}) - \min(\mathrm{a})} x\ (high - low) + low \qquad (2)$$

We computed the Pearson correlation between each groups and personality traits. If the correlation is equal or greater than 0.7, we deleted one of the pair of groups due to a highly correlated word group. A heatmap that shows the correlation between each word group is shown in Figure 1.

According to the correlation matrix, number of word groups in dataset was reduced. This correlation matrix was drawn according to the dataset with 50 tweets of each user. The initial number of attributes (word groups) was about 60. 10 of them were deleted due to including many zero values; the grammatical word groups were also deleted. After this filtering, a new correlation matrix was computed and two of them were deleted.

E. Machine Learning Models

Before evaluation of machine learning models, a statistical feature selection was employed. It is a model provided in the “Sci-kit Learn” library of Python. This function takes two parameters: a score function and the number of features that is wanted to be selected. In our project, it uses the Chi2 as a score function and returns the features with the highest scores. By applying this feature selection, only 15 of word groups that provide the highest contribution to personality trait are used in the prediction phase. Different numbers of features were used in training dataset, so we find the most successful number of features is 15 according to these tests. In order to run a classifier model, target values

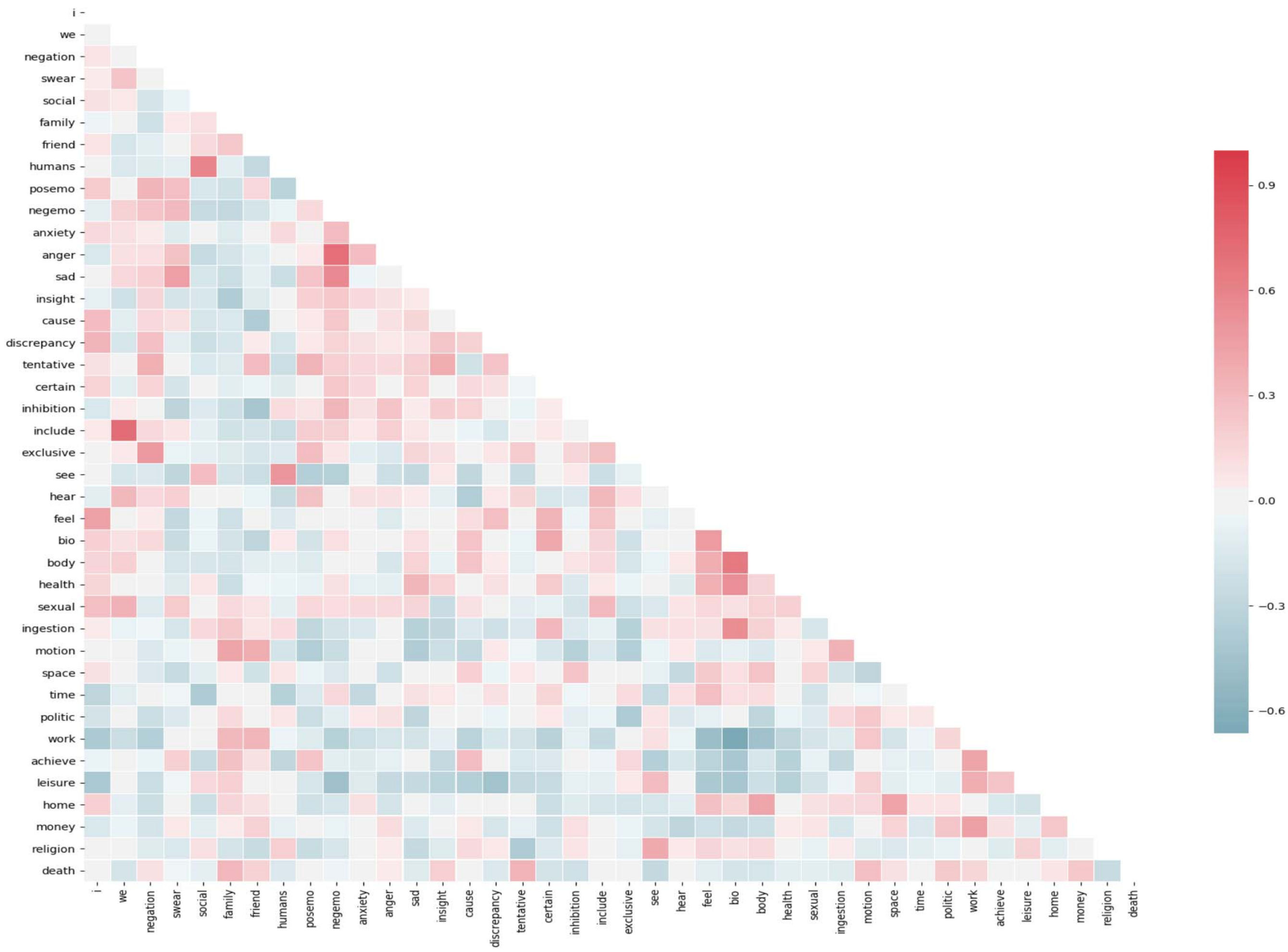

Figure 1. The Pearson correlation of each word group used in this study.

(trait scores) in the initial dataset was converted to categorical values. As a result of categorization, there are two values for the target feature: “suitable” and “not suitable”. The “suitable” category represents the personality scores in the range between [0-21] and the “not suitable” represents the scores between [22-50]. After the categorized (binary) target feature was prepared, classification models were generated. As mentioned, there are five different personality traits that are expected to be predicted. Different classification models are successful for different personality traits, since each of them has different sample data distribution. Therefore, each personality trait is predicted by using a different machine learning model. For this purpose, we experimented kNN, decision tree (DT), random forest (RF), AdaBoost, stochastic gradient descent (SGD), gradient boosting (GB) and SVM learning models. The most successful models in our experiments were AdaBoost, SGD, GB and SVM.

In the evaluation of machine learning models, a leave one out cross validation was applied due to having only 51 instances. Because there is a quite small dataset for other training methods. It basically takes only one sample as test sample and uses the rest of samples as the training data. So, each instance in dataset is used as a test data in this validation type.

IV. RESULTS

We need to decide the total number of tweets to be used in analysis before applying machine learning models. For this purpose, we prepared four different datasets. Two of them include the latest 25 and 50 tweets of users, the others include the randomly chosen 25 and 50 tweets. We did not prepare datasets more than 50 tweets, because very few people have 50 or more tweets in the dataset. To prepare the datasets with random tweets, we gathered 200 tweets of users if they have. Because Twitter allows to fetch up to 200 tweets of a user. Due to page limits, we could not give all the results of experiments. We summarize the best performing models in the below.

In the dataset with random 25 tweets; agreeableness trait has 0.761 accuracy when kNN is used, conscientiousness has 0.630 accuracy when SVM is used, extraversion 0.630 accuracy when AdaBoost is used, openness has 0.870 accuracy when SVM is used, and neuroticism has 0.783 accuracy when AdaBoost is used.

For the dataset with latest 25 tweets; agreeableness has 0.739 accuracy when RF classifier is used, conscientiousness has 0.696 accuracy when SVM classifier is used, extraversion has 0,761 when RF is used, SVM led 0.87 and 0.587 accuracy values for openness and neuroticism, respectively.

In the dataset with random 50 tweets; DT provided 0.696 and 0.652 accuracy values for agreeableness and conscientiousness, respectively; extraversion has 0.739 accuracy when SGD is used, openness has 0.870 accuracy when DT is used, and neuroticism has 0.696 accuracy when AdaBoost is used.

For the dataset with latest 50 tweets; agreeableness has 0.870 accuracy when AdaBoost is used, conscientiousness has 0.783 accuracy when SGD is used, extraversion has 0.978 accuracy when SVM is used, openness has 0.935 accuracy when GBC is used, and neuroticism has 0.761 accuracy when SGD is used.

After these experiments, the most successful models were constructed by using the latest 50 tweets of each user. Seven different machine learning algorithms were experimented. The most successful results were obtained with different machine learning models for each personality trait. As a result of using the latest 50 tweets of each user, the highest performance of 0.978 accuracy was obtained in the prediction of extraversion trait with an SVM model. Many people with the higher scores in the extraversion trait revealed a high accuracy for the prediction of this trait. On the other hand, neuroticism has the poorest prediction accuracy of 0.761 with the SGD model due to having fewer people with higher scores for the neuroticism trait. In general, the boosting and decision tree-based algorithms led the best performance in different setups.

V. CONCLUSION AND FUTURE WORKS

The ultimate goal of this study is to analyze personality with Turkish tweets and machine learning methods instead of applying only frequency-based statistical methods. Group of words that are used in previous studies for English tweets correlate between each other when they are used for Turkish tweets. The number of word groups was reduced in this study due to their pairwise correlations. To obtain successful results each personality trait is predicted by using different machine learning models.

During experimentations, we observed that the initial and the final results of models are quite different from each other, final results improved gradually. Since, when the number of tweets and users increase in corpus, the performance of models increases at the same rate.

As future works, the number of users can be increased to obtain more consistent models. The current binary classification might be replaced with more classes for a better representation of the different score levels of each personality traits, that could eventually yield more successful results. This study analyzes only tweets of a user. If other social media environments, such as blogs, Instagram etc. can be integrated for analysis, prediction results are going to be more successful than current ones. Besides word groups, different features such as number of followers, number of followings can be added to the dataset. These improvements will increase the success of prediction models.

As a conclusion, Turkish tweets can be used for predicting personality of social media users. Although we did not have a large corpus, we presented some promising results. If there would be sufficient number of instances in dataset, this study can be useful for different business areas. We proposed a pioneering approach to predict personality traits based on Turkish tweets and machine learning models, hence we believe that it would initiate new studies in this field in near future.

REFERENCES

[1] L. Qui, H. Lin, J. Ramsay, F. Yang, “You are what you tweet: personality expression and perception on Twitter”, Journal of Research in Personality, 46(6), 710-718, 2012.

[2] Y.R. Tausczik, J.W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods”, Journal of Language and Social Psychology, 29(1), 24-54, 2010.

[3] J. Golbeck, C. Robles, M. Edmondson, K. Turner, “Predicting Personality from Twitter”, IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 9-11 Oct. 2011, Boston, USA, 2011.

[4] P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, V. Sinha, “25 Tweets to Know You: A New Model to Predict Personality with Social Media”, International AAAI Conference on Web and Social Media, 16-18 May 2017, Montreal, Canada, 2017.