2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Clustering based Personality Prediction on Turkish Tweets

Esen Tutaysalgir
Department of Computer Engineering
Middle East Technical University
Ankara, Turkey
esen.aytan@ceng.metu.edu.tr

Pinar Karagoz
Department of Computer Engineering
Middle East Technical University
Ankara, Turkey
karagoz@ceng.metu.edu.tr

Ismail H. Toroslu
Department of Computer Engineering
Middle East Technical University
Ankara, Turkey
toroslu@ceng.metu.edu.tr

***Abstract*—In this paper, we present a framework for predicting the personality traits by analyzing tweets written in Turkish. The prediction model is constructed with a clustering based approach. Since the model is based on linguistic features, it is language specific. The prediction model uses features applicable to Turkish language and related to writing style of Turkish Twitter users. Our approach uses anonymous BIG5 questionnaire scores of volunteer participants as the ground truth in order to generate personality model from Twitter posts. Experiment results show that constructed model can predict personality traits of Turkish Twitter users with relatively small errors.**

***Index Terms*—personality analysis, text mining, clustering, Twitter**

I. Introduction

According to recent Twitter statistics, there are almost 400 million active Twitter users, and they are sending more than 500 million tweets each day[1]. Many companies search for ways to analyze their customers' personalities and produce customer specific advertisement strategies based on their customers' interactions on social media[2]. Similarly, politicians try to find out voter characteristics, expectations and complaints[3]. However, there are too many users, tweets, videos and photos to process. Thus, data mining techniques can facilitate dealing with such problems, and mining information from Twitter has become a popular research topic in the recent years.

Twitter has become an important tool to analyze characteristics of individuals in different fields such as psychology, marketing and politics [1]. Researchers aim to find out the personality traits from the tweets, videos, photos of the social media users. In a recent study, Querci et al. showed that there is a connection between tweets and personality traits [2]. Kwak et al. studied Twitter conversations and user relationships [3]. Zhang et al. used several features including tweet texts, hashtags, URLs, followings, retweeting relationships to investigate communities among users [4]. From social media, a variety of features can be extracted for each user. Therefore, there is a need for feature selection or reduction for such studies, especially when clustering is used for model construction [5], [6]. The Big Five personality traits is a classification for different personalities [9], which is widely accepted and used within scientific research in psychology [10]. The five personality traits, referred to as OCEAN, correspond to ***O**penness*, ***C**onscientiousness*, ***E**xtraversion*, ***A**greeableness* and ***N**euroticism*. These personality traits can be determined through personality surveys.

In this paper, our main purpose is to determine the personality traits of the users in terms of OCEAN from Turkish tweets. To this aim, we collected a limited set of ground truth data set, where the OCEAN scores are obtained through surveys matched with tweet collection of the users. We use a clustering based approach, such that the ground truth instances and the clusters they belong to are analyzed in order to obtain a prediction model. User vectors are determined through combination of extracted features and content based features. For content based features, we have utilized Turkish Part-of-Speech (PoS) tagger Zemberek[4], TF-IDF weighted term vectors, Mikolov's Word2Vec model and feature reduction with variance thresholds. After determining the best clustering schema through silhouette coefficients, the resulting clusters are evaluated with respect to the ground truth instances.

II. Proposed Method

As it is depicted in Figure 1, the proposed method includes data collection, data preprocessing, vector construction and clustering phases, which are explained in detailed below.

A. Data Collection

During data collection, firstly, we conducted a survey in order to obtain OCEAN values and the tweets of the users.

This work is supported by Ministry of Science, Industry and Technology of Turkey and Huawei Turkey with project number TEYDEB 1505-5180003

[1]http://twitter.com/twitterir

[2]https://www.neosperience.com/blog/the-new-marketing-is-people-centric-know-your-customer-personality

[3]http://abstractpolitics.com/2010/05/personality-and-political-attitudes-relationships-across-issue-domains-and-political-contexts/

[4]https://github.com/ahmetaa/zemberek-nlp

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08
http://dx.doi.org/10.1145/3341161.3343513

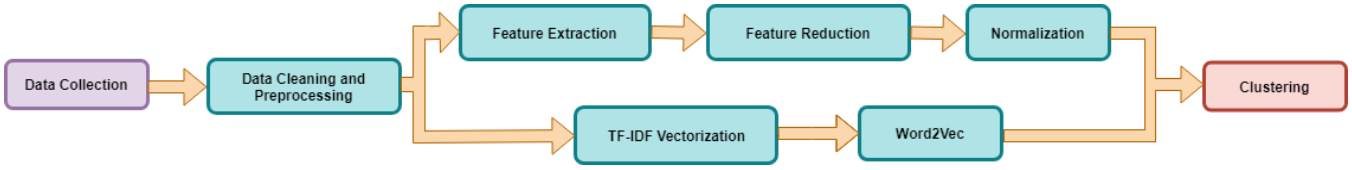

Fig. 1. Architecture of the proposed method

TABLE I
RANGE OF FEATURES

Feature	*Range*	*Feature*	*Range*
Morning	0.0 - 1.0	Past Time	0.0 - 0.88
Afternoon	0.0 - 1.0	Narrative Time	0.0 - 0.55
Evening	0.0 - 1.0	Progressive Time	0.0 - 1.5
Night	0.0 - 1.0	Condition	0.0 - 0.5
Case Ratio	0.58 - 1.0	Imperative	0.0 - 1.0
Word	3.4 - 20.0	Necessity	0.0 - 0.42
Verb	0.0 - 3.0	Ability	0.0 - 0.5
Noun	1.2 - 11.0	Negative Ability	0.0 - 0.33
Punctuation	0.0 - 5.38	Question	0.0 - 1.0
Adjective	0.0 - 2.5	Exclamation	0.0 - 0.67
Adverb	0.0 - 1.5	Ellipsis	0.0 - 0.84
Numeral	0.0 - 2.06	Full Stop	0.0 - 0.91
Determiner	0.0 - 1.0	Non-Turkish Words	0.0 - 1.0
Post Positive	0.0 - 0.726	Smiling Emoji	0.0 - 1.0
Duplicator	0.0 - 0.059	Affected Emoji	0.0 - 0.32
Conjunction	0.0 - 1.0	Tongue Emoji	0.0 - 0.16
Interjection	0.0 - 1.0	Neutral Emoji	0.0 - 0.32
Pronoun	0.0 - 1.23	Unwell Emoji	0.0 - 0.2
Question	0.0 - 1.0	Negative Emoji	0.0 - 0.06
Incorrect	0.0 - 6.58	Romantic Emoji	0.0 - 0.3
Negative	0.0 - 1.0	Fingers Emoji	0.0 - 0.56
Plural	0.0 - 3.52	Activity Emoji	0.0 - 0.036
Present Time	0.0 - 1.0	Sport Emoji	0.0 - 0.09
Future Time	0.0 - 0.5	Plant Emoji	0.0 - 0.14

About 40 volunteers participated in our survey and we have used their results as the ground truth. Afterwards, we have collected tweets in Turkish from 2000 random users by using tweepy[5] API. Then, we merged the ground truth collection with the randomly collected tweets.

B. Data Preprocessing

Before constructing the user vectors, the following preprocessing tasks have been applied on the collected tweets.

Data Cleaning: As conventionally applied for tweets, mentions, hashtag symbol "#", *RT* keyword, and URLs are removed from the texts. Hashtag keyword is kept since it may include a meaningful phrase related to the tweet or to the user. If all tweets of a user are composed of *RT* keyword or URLs only, the user is neglected. Additionally, Turkish stopwords[6] are removed from tweets in order to improve accuracy for tf-idf weighting and word2vec embedding.

Text Normalization: Due to informal language used in Twitter, as conventionally applied before NLP operations on the text, content of the tweets are normalized. To this aim, we have used Zemberek's SpellChecker tool, such that incorrect words are detected and replaced with the best alternatives. During this normalization process, deasciification is applied as well. If a tweet is written with non-Turkish characters, words are corrected with corresponding Turkish characters by Zemeberek's deasciification tool.

Lemmatization: As frequently applied in text mining, we replaced the words in tweets with their lemmas so that the words with different forms are represented as the same for tf-idf weighting and word2vec embedding.

C. Vector Construction

As seen in Figure 1, after data cleaning and preprocessing, we applied two different paths for vector construction, and then combined the resulting features. As the first path, we extracted features from the use of language and timestamp of the tweet. As the second path, terms within the texts are used as the features.

1) Feature Extraction: As listed in Table I, we have constructed 48 features. Most of them are related to the use of non-Turkish words, emoticons, numerals and punctuation, etc. in tweets. The weight of the features is determined with the average frequency observed within all tweets of a user. The temporal features extracted from tweet timestamp require extra treatment. Tweet timestamps are discretized as *Morning*, *Afternoon*, *Evening* and *Night*, where each one is defined as a separate feature. Since these time intervals are circular, one-hot encoding is not suitable to calculate average tweet time and to obtain distance between time values. Hence, we employed two-hot encoder as follows:

- Tweet time is *Morning* →Add 1 to *Morning* and *Night*
- Tweet time is *Afternoon* →Add 1 to *Morning* and *Afternoon*
- Tweet time is *Evening* →Add 1 to *Afternoon* and *Evening*
- Tweet time is *Night* →Add 1 to *Evening* and *Night*

Then, the average values for each temporal feature is calculated. In this way, euclidean distance can be used for as the distance measure between temporal features.

Another important set of features is obtained through *Part of Speech (PoS) tags* used within the texts. To this aim, we determine PoS of the words in the tweets through Zemberek, and average frequency of each PoS tag is recorded as the weight of the feature.

2) Feature Reduction: Before clustering, we have applied feature reduction as follows: As given in Figure 2, we have plotted the sorted distribution of the values for each feature

[5]https://www.tweepy.org/
[6]https://www.ranks.nl/stopwords/turkish

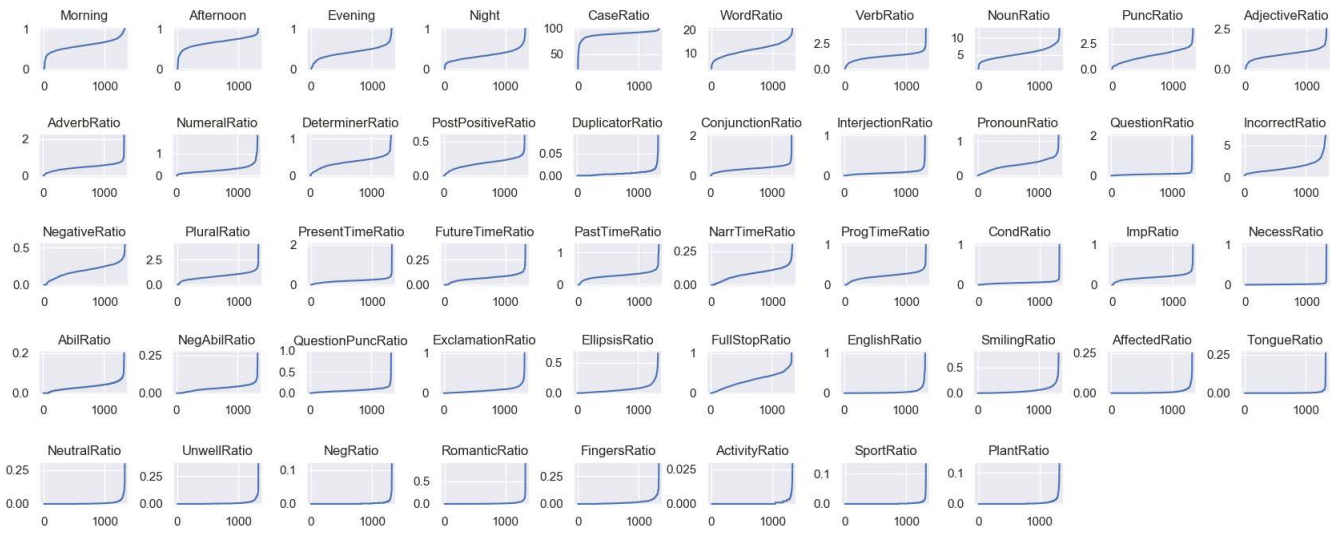

Fig. 2. All sorted feature values for all users

for 1000 users. As seen in the plots, values for some of the features are nearly the same for most of the users, hinting that these features are not discriminative for clustering. Hence, we filtered the features according to a variance threshold. We calculated variances for all features and selected our threshold value as 0.01. Hence, the features having a low-variance (i.e., lower than 0.01) are removed. After feature reduction, the following 20 distinguishing features are retained: *Morning*, *Afternoon*, *Evening*, *Night*, *Word*, *Verb*, *Noun*, *Punctuation*, *Adjective*, *Adverb*, *Negative*, *Numeral*, *Determiner*, *Conjunction*, *Pronoun*, *Incorrect*, *Plural*, *Full Stop*, *Smiling Emoji*, *Negative Emoji*.

3) Normalization: We have applied different normalization techniques on our feature vectors, which are standard scaling, robust scaling and discretization. The effect of each normalization technique on the clustering performance is analyzed in order to determine the best one for our problem.

4) TF-IDF Weighting and Word2Vec based Word Embedding: As the second path of vector construction, we used the terms in the tweets of a user. As the first step, in order to find the most important words or phrases used in the user's all crawled tweets, we calculate TF-IDF values for 1-gram, 2-grams and 3-grams. Then, Word2Vec embeddings of the top words (i.e., with the highest TF-IDF values) are constructed. Finally, user vector with 38 dimensions is constructed through concatenation of Word2Vec representations of the top words.

5) Composition of Extracted Features and Word2Vec Vectors: After obtaining features on both paths, i.e., extracted features and Word2Vec representations of the top terms in the tweets, the two sets of features are concatenated to construct the final vector. As the result of this step, we have obtained a matrix with 58 columns for 1923 users.

D. Clustering

As explained in the previous section, we have applied a standard survey to obtain OCEAN scores for users. As the result of the survey, the OCEAN dimensions are scored in terms of percentage. For each dimension, we have discretized the scores into 4 blocks, as 0-25%, 25-50%, 50-75% and 75-100%. Thus, for 5 OCEAN scores, there are 20 maximum number of possibilities in order to cluster the users. In order to find groupings, we applied two clustering algorithms, K-means and Agglomerative clustering. Clustering quality is measured with silhouette-coefficient. We used scikit-learn[7] library for the implementation of the methods. In k-means algorithm, k-values are selected as 4, 8, 12, 16 and 20, to determine how silhouette coefficient is changing with the k value. In order to accelarate the convergence, *k-means++* initialization method in scikit-learn library is used. For similarity calculation, euclidean distance metric is applied. Agglomerative clustering is applied under 4, 8, 12, 16 and 20 clusters. The effect of different linkage options, including single, average and complete linkage, is tested, and the best result is obtained with average linkage. In addition, we used euclidean distance as the distance metric.

III. Experiments

A. Error Rate Calculation

The user vectors and discretized OCEAN scores of 40 users are used as the ground truth in order to calculate the error rates of the clustering results. In the rest of the paper, we call these survey users as the *base users*. To this aim, we have used the following simple approach: Firstly, we found the highest k value (i.e., the number of clusters) such that each base user has a base neighbor in the same cluster. In other words, if there is only one base user in a cluster, then we backtracked to the previous k value for clustering. We calculate the error rates for each OCEAN dimension separately as follows: we consider the base users in the same cluster, and for each user, we calculate

[7]https://scikit-learn.org

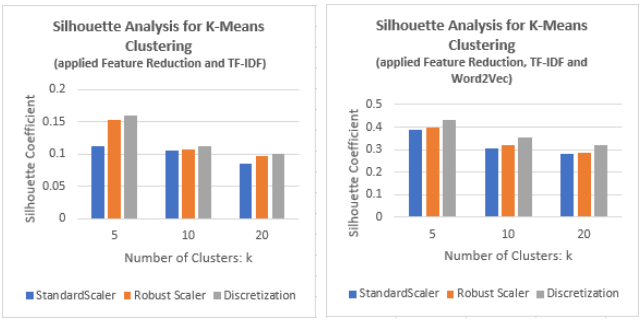

Fig. 3. Clustering quality results under k-means clustering without and with Word2Vec embedding

the average score of its neighbors. Then, the average value is subtracted from the selected user's score. The result is the error rate for that base user. Then, we have repeated the process for all base users in the same cluster. Base error rate of a cluster, for the corresponding OCEAN dimension, is calculated as the average error rate of all base users in the cluster.

B. Experiments

We have performed 4 different experiments. In the first experiment, only TF-IDF weighting is applied to the user feature vectors. Moreover, we have applied different normalization techniques on the extracted feature vectors and combined it with the TF-IDF vectors. Then, k-means is used as the clustering algorithm. For this experiment, silhouette coefficient around 0 has been obtained. In the second experiment, firstly, TF-IDF weighting schema is applied, and then, the resulting vectors are given to Word2Vec embedding as the input. After that, k-means is used as the clustering algorithm. We have obtained higher silhouette coefficient scores than that of the the first experiment. The results for these experiments are given in Figure 3. In the third and the forth experiments, agglomerative clustering is used instead of k-means clustering. For these experiments, as shown in Figure 4, similar silhouette coefficient scores are obtained. Since k-means clustering with Word2Vec embedding provides the best silhouette coefficient scores, we can consider this as the basis prediction model. The error rates for each dimension of OCEAN with this model are given in Figure 5.

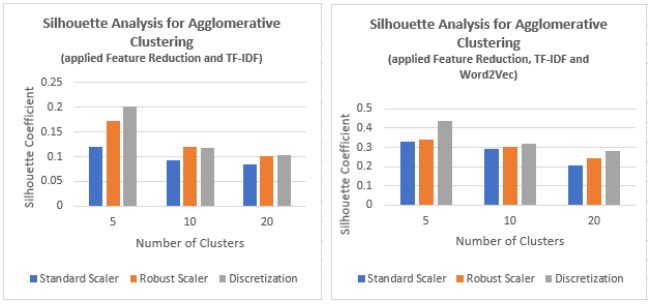

Fig. 4. Clustering quality results under agglomerative clustering without and with Word2Vec embedding

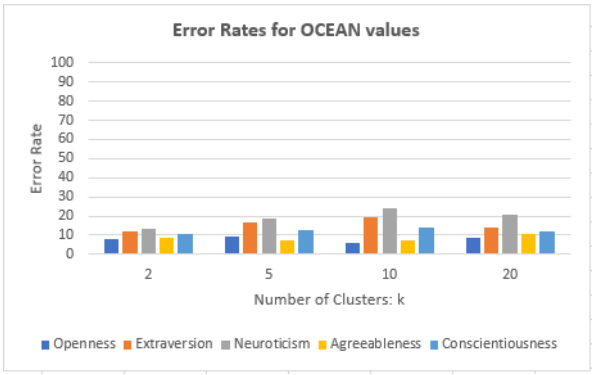

Fig. 5. Error Rates

IV. Conclusions

In this paper, we present a clustering based framework for predicting personality traits of users through their tweets. The framework has been designed especially for Turkish tweets. The process starts with data cleaning and preparation step, followed by determining tweet related and language use related features from the tweets. Afterwards, feature reduction is applied Word2Vec embedding of the remaining terms are generated. Due to the relatively small number of ground truth personality trait values obtained, we devised a error measurement mechanism based on how well the ground truth instances are clustered. The experiments reveal that the overall approach produces promising results, and it is open to further improvements.

References

[1] J. Bollen, H. Mao, and A. Pepe. "Determining the public mood state by analysis of microblogging posts", In Procceedings Of the Alife XII Conf. MIT Press, 2010.

[2] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter", 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011.

[3] H.Kwak, C. Lee, H. Park, and S.Moon, "What is Twitter, a social network or a news media?" in Proceedings of the 19th International World Wide Web Conference (WWW 10), pp. 591600, 2010.

[4] Y. Zhang, Y. Wu, Q. Yang, "Community Discovery in Twitter Based on User Interests", Journal of Computational Information Systems, 2012

[5] V. Roth, T. Lange, "Feature Selection in Clustering Problems", Advances in Neural Information Processing Systems 16, 2004.

[6] M. H. C. Law, M. A. T. Figueiredo and A. K. Jain, "Simultaneous feature selection and clustering using mixture models", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 9, pp. 1154-1166, 2004.

[7] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Fifth International AAAI Conference on Weblogs and Social Medi, 2011

[8] S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter", In Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010

[9] Barrick, M. R. and Mount, M. K., "The Big Five Personality Dimensions Aand Job Performance: A MetaAnalysis," Personnel Psychology, 44: 1-26., 1991

[10] L. Goldberg, J. Johnson, H. Eber, R. Hogan, M. Ashton, R. Cloninger, and H. Gough. "The international personality item pool and the future of public-domain personality measures." Journal of Research in Personality, 2006.