ParsTSet: A Persian Dataset for Personality Prediction on Twitter

Mohammad Mahdi Abdollahpour*, Zahra Anvarian* , Samin Fatehi , Sauleh Eetemadi
School of Computer Engineering, Iran University of Science and Technology, Iran
{mohammadmahdiabdollahpour, zahra.anvarian97}@gmail.com,
sa_fatehi@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

In recent years, recognizing individuals' personality traits through social media has become an interesting topic in both fields of natural language processing and social sciences. Psychological research also shows that some personality traits are associated with language behavior. NLP models can take advantage of this correlation to model and predict personality traits based on the vast amount of data available, thanks to modern social media. No such dataset exist for the Persian language. We have constructed a novel dataset labeled with Myers-Briggs Type Indicators (MBTI) consisting of 1,552,532 tweets. We present our data collection method and discuss its challenges and results in detail. We also introduce a baseline classification model by fine-tuning a variation of BERT architecture (ParsBERT), pretrained on Persian corpora.

1 Introduction and Related Work

Research in personality detection can be helpful in a variety of other fields, including job screening (Liem et al., 2018), recommendation systems (Yang and Huang, 2019), advertising (Matz et al., 2017), word polarity detection (Poria et al., 2019), and social network analysis (Balmaceda et al., 2014).

The Myers-Briggs psychological model (MBTI) refers to patterns of how the world is viewed, information is collected, how decisions are made, and how individuals live out lifestyle choices (Martin, 1997). There are four continuous traits in the MBTI model (Myers, 1962):

- Extroversion (E) vs. Introversion (I)
- Sensing (S) vs. Intuition (N)
- Thinking (T) vs. Feeling (F)
- Judging (J) vs. Perceiving (P)

Since people's writings represent their identity, their writings can be used to detect their personality (Mehta et al., 2019). Several datasets have been collected and analyzed for personality prediction in English and some other languages (Gjurković et al., 2020; Mehta et al., 2020; Amirhosseini and Kazemian, 2020; Kazameini et al., 2020; Lynn et al., 2020; Gjurković and Šnajder, 2018; Tandera et al., 2017; Majumder et al., 2017; Plank and Hovy, 2015a; Kosinski et al., 2013; Pennebaker and King, 1999), and one is present openly on Kaggle[1]. However, no dataset has been prepared in this field for the Persian language. By searching for identifier phrases and preparing a questionnaire, we have collected a dataset composed of more than 1.5M tweets written by 938 individuals in Persian labeled by their MBTI personality type.

2 Dataset Construction

The dataset is comprised of two main parts: First, a collection of publicly available tweets (inputs), and second, each individual's personality traits (labels). Two primary techniques could help to acquire each user's personality traits. The first one is to search for keywords or sentences indicating any personality trait (Plank and Hovy, 2015b). In our case, it would be searching for MBTI keywords[2] in the Bio section and Tweets. For example, a keyword like "INFP" or a sentence similar to "I am an ESTP", translated to Persian, of course. This technique enabled us to collect more than 92% of our data. The second technique was to request people via various channels to fill out a questionnaire asking for their Twitter handle and their MBTI personality class if they already know it. Since many people are not familiar with this test, we have included instructions and links to help them conduct the test and

[1] kaggle.com/datasnaek/mbti-type

[2] According to the MBTI model, there are 16 keywords as a result of concatenating abbreviations of each trait. For instance: I+N+T+J=INTJ. They are written exactly in the same spelling in Persian.

	Method I		Method II	
	Bio	Tweet	Questionnaire	Total
# Tweets	309,364	1,134,294	108,874	1,552,532
# Users	210	653	75	938

Table 1: Detailed volume of the collected data

continue filling out the questionnaire afterwards. We distributed the questionnaire on many different online channels, but in terms of validity and the amount of data, the best results were obtained via the responses from Twitter itself.[3] Also, we desgined a branching logic for the questionnaire using Microsoft Forms that successfully reduced the invalid submission rate. Finally, we conducted an automated sanity check on $\frac{1}{3}$ of the data.

2.1 Methods Comparison

The first method helped us collect 1,443,658 tweets in contrast with the second method that only contributed to 8 percent of the data (Table 1). From our point of view, there are several reasons for this disparity. First, convincing people to fill out a questionnaire is a tedious task. To overcome this challenge, some teams use motivational and sometimes forcing techniques that are impractical or unethical.[4] Second, finding a target group who are interested or knowledgeable about a topic is challenging. For this purpose, one idea would be to ask users with a relatively high number of active followers to redistribute our questionnaire, which we tried with adverse feedback. However, in comparison with the first one, this method has the benefit of not needing to intervene constantly, hence being able to make progress in parallel.

3 Data Analysis

In Iran, with a population of more than 82M, 62,519 people took the MBTI test on a well-known website[5]. We made a comparison between our dataset and the data from that website, as an estimation of Iran's MBTI statistics, which is depicted in Figure 1. We observe that introverts are less present in society than in social networks. This is because these people are better able to socialize and be active in social media and more openly express their feelings (Goby, 2006).

[3] All respondents consented to the usage of their data.
[4] No individual has been incentivized in this research.
[5] 16personalities.com/country-profiles/iran

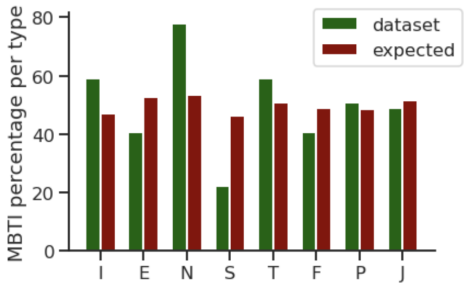

Figure 1: Distribution of each attribute in our dataset compared with its estimated distribution in Iran

We re-formatted data and fine-tuned ParsBERT (Farahani et al., 2020), a pre-trained BERT-based model (Devlin et al., 2018), and a logistic regression classifier on top of [CLS] encodings separately for every trait. Note that when splitting the data, all tweets of each user have to be in only one set. Before feeding the data into the model, we balanced every category by downsampling the larger categories. We "evaluated" the model using repeated stratified K-fold (with k, n = 5) technique. To be more clear, being in the first steps of the research, we did not intend to tune hyper-parameters or go through a model selection process; thus, we did not need to use nested K-fold.

We got 56.96 for the mean of macro average F1-scores over results of all iterations over four traits. Being an unfavorable result, there is not much value in reporting detailed metrics. However, this is an ongoing research, and we plan to improve the results both in terms modeling and data quality. We attained encouraging results via our enhanced data and models which will be presented later.

4 Conclusions and Future Works

In this paper, we have demonstrated the first steps in building a Persian dataset for personality prediction. We trained a basic classifier as a baseline. In the future, we will collect more data, experiment with more sophisticated models, and team up with psychology experts to evaluate and improve our dataset. We even achieved promising early results using more complex classifiers over fastText (Bojanowski et al., 2017) and doc2vec (Le and Mikolov, 2014) embeddings using a different training approach. Furthermore, overcoming the challenges of gathering data using a more science-backed personality test, BigFive, is a gateway to further improvements in modeling psycholinguistic features of a text in Persian.

References

Mohammad Hossein Amirhosseini and Hassan Kazemian. 2020. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, 4(1):9.

Jose Maria Balmaceda, Silvia Schiaffino, and Daniela Godoy. 2014. How do personality traits affect communication among users in online social networks? *Online Information Review*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *arXiv preprint arXiv:2005.12515*.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2020. Pandora talks: Personality and demographics on reddit. *arXiv preprint arXiv:2004.04460*.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97.

Valerie Goby. 2006. Personality and online/offline choices: Mbti profiles and favored communication modes in a singapore study. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 9:5–13.

Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. 2020. Personality trait detection using bagged svm over bert word embedding ensembles. In *Proceedings of the ACL 2020 workshop on Widening NLP*. Association for Computational Linguistics.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning*, pages 197–253. Springer.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Charles R Martin. 1997. *Looking at type: The fundamentals*. Center for Applications of Psychological Type.

Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

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

Barbara Plank and Dirk Hovy. 2015a. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Barbara Plank and Dirk Hovy. 2015b. Personality traits on twitter—or—how to get 1,500 personality tests in a week. pages 92–98.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. 2017. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611.

Hsin-Chang Yang and Zi-Rui Huang. 2019. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems*, 165:157–168.