Analysis of Big Five Personality Traits by Processing of Social Media Users Activity Features

© Maxim Stankevich © Ivan Smirnov

Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

RUDN University, Moscow, Russia

stankevich@isa.ru ivs@isa.ru

© Nikolay Ignatiev

RUDN University, Moscow, Russia

naignatiev@yandex.com

© Oleg Grigoriev

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

oleggpolikvart@yandex.ru

© Natalia Kiselnikova

Psychological Institute of Russian Academy of Education, Moscow, Russia

nv.pirao@gmail.com

Abstract. The study focused on the analysis of relation between Big Five personality traits of a user and his activity in popular Russian social media Vkontakte. In order to receive Big Five personality trait scores, we asked Vkontakte users to complete a psychological survey and then analyzed data from their personal public social media pages. The purpose of the study was to investigate the relation between social media activity features and users' level of neuroticism, conscientiousness, extraversion, openness to experience and agreeableness. To perform the task, we used machine learning classification algorithms.

Keywords: social media analysis, big five personality traits, machine learning, classification.

1 Introduction

The Big Five personality traits model is a popular psychological tool, which is commonly used for describing the human personality through the following measurements: neuroticism, conscientiousness, extraversion, openness to experience and agreeableness [1]. Personality traits scores are usually calculated with the help of questionnaires. Widespread use of social media makes it possible to receive information about social media users by analyzing data retrieved from their public pages. However, there are only a few studies related to the analysis of users' Big Five personality traits by using social media activity information from Russian-speaking social networks. A number of researchers are involved in Big Five personality traits prediction and analysis for English-speaking social networks [2,3], but there are no in-depth studies for Russian. The proposed approach and the dataset thus collected are new for the Russian social network analysis. The purpose of the study was to investigate the relation between social media activity features and user's level of neuroticism, conscientiousness, extraversion, openness to experience and agreeableness by using machine learning algorithms.

In order to form the dataset, we asked volunteers to complete NEO-FFI questionnaire [4] and then to provide access to their public pages information under privacy constraints. Thus, we received data of 165 users from popular Russian social network Vkontakte. We presented five personality traits scores on the following scale: low level, medium, and high. The idea is to present the problem as multiclass classification. Classification features are based on a 1-year period of user activity represented as posts on their public pages and general information about users' profiles such as gender and a total number of friends and followers. To evaluate methods, we ran two sets of experiments with different classifiers: support vector machine and random forest.

The main issue that we faced with was a lack of training examples. Though because of insufficient data we couldn't significantly improve classification performance, we came to the conclusion that feature format should be redesigned and text analysis-based features should be added. We continue data collection and look forward to improve our results in the nearest future.

2 Related works

There is a lot of studies that investigate social media-based data usage for classification in different psychology related tasks.

Besides Big Five personality analysis, detection of depression, post-traumatic stress disorder and anxiety, is also a very important problem. For example, CLPsych 2015 Shared Task organizers built the dataset consisting of messages collections of depressed and non-depressed users and asked contributors to share the performance of their depression detection models [5]. This shared task as well as the other similar studies, such as [6] and [7] used textual data natural language processing methods to form features for a predictive model. Authors of related works [8] and [9] used a social media activity features to improve classification performance. One should take into account that depression detection task is time-dependent – it is necessary to consider time constraints while dataset preparation, at the same time Big Five personality traits are more consistent in time [10].

One of the most significant studies, related to social media language and Big Five personality traits, presented in [3]. The authors performed analysis of 700 million words, phrases, and topic instances collected from Facebook messages of 75,000 volunteers, who took a standard personality test. The work demonstrates some important dependencies between language use and users' personality attributes. For each personality trait, they formed the list of related words which showed valuable correlations with a neuroticism, conscientiousness, extraversion, openness to experience and agreeableness levels.

The work presented in [11] describes the Big Five personality traits prediction models for Twitter users. This research dataset contains most recent 2000 tweets of 279 volunteers. To perform the task authors decided to present personality traits scores as values on a normalized 0-1 scale. The features used were based on text analysis. The authors utilize Linguistic Inquiry and Word count tool [12] to produce statistics on 81 different features. The MRC Psycholinguistic Database [13] was used to retrieve features from users' vocabulary. The authors also performed social media activity features. The results of correlation analysis revealed that some of them had correlations with five-factor personality model. The proposed models showed about 15% of mean absolute error on a normalized scale for each personality trait as a measurement of model prediction accuracy.

Another work related to the task of Big Five traits prediction described in [14]. The data for the research contains information about *likes* of 58,466 volunteers from the Facebook social media. The authors used decomposed *User-Likes* matrix with logistic and linear regression classificators to predict users Big Five personality traits and other

personal attributes. The model achieved high prediction performance for personal attributes such as gender, age, and nationality (~80% Area Under Curve) and about 35% of accuracy score for users' Big Five personality traits.

The research presented in [15] has a similar with our work idea. Authors collected users' data from Vkontakte and performed correlation analysis using social media activity indicators. The main interest was on photos published on the users' public pages. According to the results, most significant correlations were found between extraversion and such activity indicators as a number of friends and followers, total numbers of posts and some photo information-based indicators. Neuroticism score also showed valuable positive correlation with a users' total number of posts.

We analyzed related works and came up to the following conclusion. The background studies propose valuable methodologies for Big Five personality traits analysis and prediction which are mainly related to language use of English-speaking social media users. For Russian-speaking social networks this problem is not well studied. For example, in 2007 myPersonality project[6] started to gather social media data and results of psychology questionnaires from Facebook users. The huge volume of this project was successfully used for different academic studies. However, there are no available and appropriate datasets based on Russian-speaking social media. This is the main reason why we had to form our original for the task of Big Five personality traits analysis of Vkontakte users.

3 Dataset

To build the dataset we asked volunteers from Vkontakte to take part in a psychological survey and complete NEO-FFI questionnaire. After this part, we requested access to their public pages under privacy constraints. Finally, for those who provided their acceptance and completed questionnaire we collected all available information from their public profile pages. Overall, data from 165 profiles was assembled. Personal information that can reveal the identity of a persons was removed from the data.

We divided collected data into two categories: general information about users and information about user messages posted during the time period from January 2017. The first part contains such features as - number of friends, number of followers, gender, number of followed groups and communities, etc. The second part contains the text of the users' messages, timestamps, and numbers of *likes*, *commentaries,* and *reposts* (analog of a retweet on Twitter).

It is worth mentioning, that we continue to expand our dataset with new examples. This study is based on

---

the current amount of available data, but we consider this number only as an intermediate stage.

4 Methods

4.1 Big five personality traits

Here we describe the methodology for the Big Five personality traits prediction. We also describe the personality scores representation and features that we extract from available data.

As a first step, we divide the initial NEO-FFI score scale (0-48) of each personality trait as following: *low level* (0-20), *medium level* (21-32) and *high level* (33-48) [16]. As a result, one of these three classes were assigned to each of user's scores of neuroticisms, conscientiousness, extraversion, openness to experience and agreeableness. Thus, the initial task is transformed to the task of multiclass classification. It should be noted that such approach imposes some restrictions on evaluation method. Figure 1 represents the class distribution among users' level of extraversion.
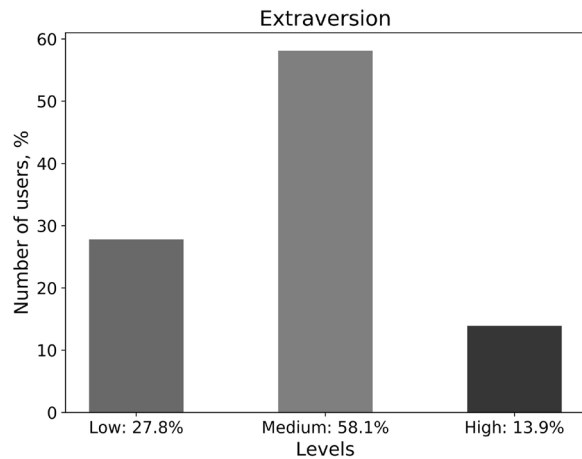
Figure 1 Levels distribution for users' extraversion score.

Despite the fact that the *medium level* covers the shortest score interval, Figure 1 illustrates that the majority of users fall into this class. The same situation is observed with other personality traits. The statistics for each of Big Five personality trait presented in Table 1.

4.1 Features

The format of Vkontakte personal page provides a wide range of user information. We used gender, number of friends, number of followers, number of followed groups, number of photo, and number of audio tracks to form a users' feature set. While filling out a Vkontakte personal page, users can provide their opinion on predefined question such as, how they relate to smoking or what is the most important in people and life. Our data include all the answers, but it is hard to present such information as a feature. Since these questions are not mandatory for Vkontakte users, we decided to assign

Table 1 Big Five personality traits label distribution among users in the dataset.

Neuroticism, %	Low:	33.9
	Medium:	49.6
	High:	16.3
Conscientiousness, %	Low:	19.3
	Medium:	55.7
	High:	24.8
Extraversion, %	Low:	27.8
	Medium:	58.1
	High:	13.9
Openness to experience, %	Low:	10.6
	Medium:	66.3
	High:	23.1
Agreeableness, %	Low:	15.1
	Medium:	73.3
	High:	11.5

them binary values that represent if a user provided this information or not. We assume that these answers can indicate users' general readiness to share their opinion with other people and that might be valuable for future analysis.

As it was mentioned, we collected users' messages from their public pages. We used information about likes, commentaries and reposts related to these messages to calculate their averaged values on a single post. The fact that for every user we collected messages posted during an equal time period allows as to use total number of assembled posts as a feature. The messages timestamps were used to calculate the proportion of users' messages posted during night time (12 P.M – 6 A.M.).
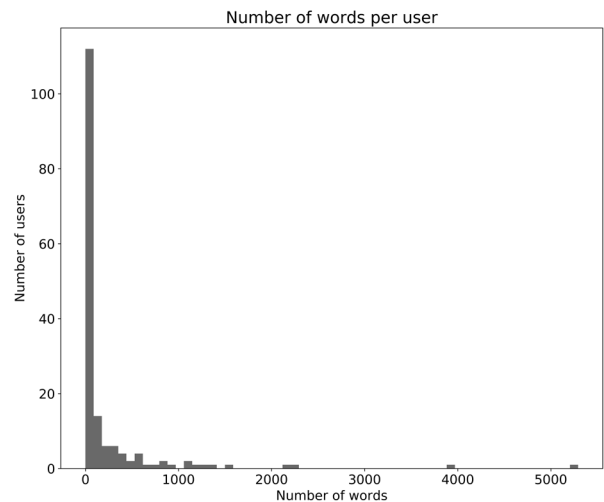
Figure 2 Number of words in users' messages.

Table 2 Averaged results of multiple 4-fold cross-validation runs on the data.

Random Forest			
Big Five trait	*Recall, %*	*Precision, %*	*F1-score, %*
Neuroticism	**49.07**	**53.01**	**49.51**
Conscientiousness	35.19	37.12	35.46
Extraversion	**46.41**	46.79	**46.38**
Openness to experience	**44.65**	47.50	**45.46**
Agreeableness	**51.15**	56.04	**53.15**
SVM			
Big Five trait	*Recall, %*	*Precision, %*	*F1-score, %*
Neuroticism	33.88	49.17	33.38
Conscientiousness	**37.54**	**41.69**	**36.17**
Extraversion	40.02	**47.04**	41.78
Openness to experience	32.28	**52.47**	35.07
Agreeableness	38.10	**57.59**	43.26

However, Vkontakte profiles in personal pages provide much less text data than Facebook and Tweeter. The most popular format of Vkontakte users' activity is reposting. A large amount of communities provides different kind of content and users usually only repost this content on their personal pages without giving any commentaries or opinions. Overall, we collected 13152 posts, but majority of them were empty reposts. Only 2637 of them contain texts written by users themselves. The total number of used words for each user is presented on Figure 2.

As we can see on the Figure 2, current data contains a very limited amount of information about Vkontakte language. Considering this, we decided to perform classification without language analysis. It is necessary to collect much more data before applying text analysis and compiling text-based features. In this work, we perform classification task using mostly social media activity features.

Despite this fact that we ignored lexical features in this research, we processed messages data to form several additional features. For example, the average number of sentences and words. We also computed the proportion of uppercase words as well as the number of ellipses in the users' writings. We assume that described features could reveal some specifics of people's behavior in social media.

5 Results of experiments

The following chapter represents the results of our experiments. To perform the evaluations, we used scikit-learn implementation of random forest and multiclass SVM algorithms [17]. The parameters for the classification were set up by grid-search with 4-fold cross-validation.

We calculated the macro variation of recall, precision, and f1-score to present classification performance. To evaluate the accuracy of our models we compiled 10 runs of 4-fold cross-validation on the data. The results of our experiments presented as an averaged value of these runs for each metric. The multiclass classification results with a 4-fold cross-validation presented in Table 2. The best values for each metric highlighted in bold.

The best performance was shown for the agreeableness and neuroticism with a 49% and 53% of f1-score respectively. The slightly worse results were received for extraversion and openness to experience with a 45% and 46% of f1-score. Random forest classification algorithm was used to get these results. The conscientiousness personal trait performance was the lowest in our experiments with only 36% of f1-score received by SVM. It is worth to note that in the most cases SMV achieved more precision than RF, but recall score was significantly less.

In general, we can't define considered performance as good. However, limited information about language use of Vkontakte users prevented the possibility to compile lexical features and perform text analysis. According to the results of studies based on English-speaking social media, text features might serve as an effective revealing tool for users Big Five personality traits. Thus, in this study, we mostly tested social media activity features, which we can describe as being useful for the considered task.

6 Conclusion

In this work, we performed the prediction of Big Five personality traits of social media users. We collected results of NEO-FFI questionnaire taken by 165 volunteers and compiled dataset using social media activity information from their personal pages. The

personality traits scores were represented as *low*, *medium*, and *high* levels to transform the task into multiclass classification.

We can define two limitations that we faced during our work. The first one consists of the fact that Vkontakte users' messages provide a very small amount of text data. We observed that collected messages, for the most part, are empty reposts, which don't provide any text written by users personally. This limitation imposes some restriction on our current study. The features for the classification were compiled by processing of social media activity information without any lexical features. We assume that such features can greatly improve classification results. The second limitation is a simple lack of examples in our current dataset.

Considering this limitation, we can admit that our most important task now is to add much more new examples to the dataset. With a greater size of data, we can utilize text analysis approaches and investigate the relation between Big Five personality traits and Russian-speaking social media language, which is currently an unresearched field of study.

References

[1] Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in personality, 37(6), 504-528.

[2] Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. Journal of computer and System Sciences, 80(1), 57-71.

[3] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

[4] Costa, P. T., & McCrae, R. R. (1989). NEO five-factor inventory (NEO-FFI). Odessa, FL: Psychological Assessment Resources.

[5] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 31-39).

[6] Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., ... & Sheth, A. (2017, July). Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 1191-1198). ACM.

[7] Jamil, Z. (2017). Monitoring Tweets for Depression to Detect At-risk Users (Doctoral dissertation, Université d'Ottawa/University of Ottawa).

[8] De Choudhury, M., Counts, S., & Horvitz, E. (2013, May). Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 47-56). ACM.

[9] Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013, April). A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 201-213). Springer, Berlin, Heidelberg.

[10] Cobb-Clark, D. A., & Schurer, S. (2012). The stability of big-five personality traits. Economics Letters, 115(1), 11-15.

[11] Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011, October). Predicting personality from twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 149-156). IEEE.

[12] Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.

[13] Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497-505.

[14] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

[15] Shchebetenko, A. (2013). Big Five and usage of the VK online social network. Bulletin of South Ural State University, Series "Psychology" (pp. 73-83).

[16] Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. Psychological assessment, 4(1), 5.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830