Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210207

Social Media User Profiling Based on Genre Extraction

Konstantin BELOUSOV and Ivan LABUTIN [1]
Perm State University, Perm, Russia

Abstract. The task of automatic user profiling (in particular, determining their psychological parameters from their texts) in Social Networking Services (SNS) is of great practical importance in many fields (PR and marketing, advertising, politics, social relations and recommendations, etc.). However, this problems' solution is often complicated by the need to process large amounts of data and the inability to explain the results achieved. Our article presents a new extensible fuzzy classification method for social media user profiling based on preliminary expert analysis of the linguistic behavior of such users. The proposed method is akin to topic modelling, but is not computationally expensive (so it can be used for large-scale data / web text analysis) and produces results that are relatively easy to interpret. Comparison with the other methods presented in literature also testifies in favor of the approach. The profiling accuracy reaches 65-70% on a relatively small dataset for such kind of studies.

Keywords. Fuzzy Classification, Text Mining, Information Extraction, User Profiling, Social Media, Behavioural Analysis

1. Introduction

Existing works on the user behavior profiling of social network services (SNS) are based on the analysis of social, psychological and behavioral (social circle, likes, the amount of time spent on the network, etc.) aspects (see [1,2]), while the linguistic component of the communication process is not considered. We apply the approach of the unity of behavior (including that determined by social and psychological characteristics) and cognition that exists in the cognitive sciences [3,4]. With regard to our context, we can talk about the relationship between the psychological characteristics of a person and the parameters of his speech behavior in a social media. Confirmation of this hypothesis would open up the possibility of auto-tagging content and profiling of Social Networking Services users.

This formulation of the hypothesis poses the problem of analyzing text arrays consisting of user-generated content from social media. The problem is solved by using the information system of graphosemantic modeling "Semograph" [5] developed by the team. The cornerstone of "Semograph" is a concept of *semantic fields* – sets of linguistic units (words, phrases and even more compound constructions) with meanings that share common semantic feature. Semantic fields can comprehensively describe textual and hy-

[1]Corresponding Author: Ivan Labutin, Perm State University, Bukireva Str. 15, 614990 Perm, Russia; E-mail: i.a.labutin@yandex.ru.

pertextual content in terms of semantics, syntax, graphics (for example, the use of special punctuation or emoticons), stylistics (for example, message genres [2]), etc. This complex description makes it possible to use the sets of semantic fields as the representations of contexts that have been marked up with these fields.

In this work, our task was to implement methods for determining the psychological parameters of a person based on the expert analysis [3] of their speech behavior in social media performed in the IS "Semograph". This article is a continuation of our research [6, 7] and presents the methods we used to solve this problem for users of Russian Social Networking Services.

The theoretical significance of the study lies in determining the dependencies between non-speech and speech parameters of behavior on the one hand, and behavior and cognition on the other, which, ultimately, can be used for automatic content tagging and user profiling of social media services.

From the practical standpoint, the distinctive features of the proposed method are low computational cost and the inherent explainability coming from its linguistic nature.

2. Related Work

A review of works on the problem of studying SNS indicates that several main methodological approaches to the study of the personality of social media users are formed in linguo-personolology [8]. The following main groups of approaches can be distinguished:

- Building a set of user characteristics based on thematic and psycholinguistic dictionaries (LIWC [9], MRC [10]) [11]. However, such kind of dictionaries in the open access for the Russian language is almost non-existent – there is Russian LIWC spin-off, but it is far from being as complete as the English one.
- Selecting behavioral characteristics as parameters for the profiling models, such as "likes" under images of certain brands, friendship connections, geotags in photos [12]. But such data is usually quite difficult to extract from a social media sources and it usually requires heavy preprocessing.
- Building the parameters of a user profiling model directly from texts using machine learning – for example, using neural networks together with Word2Vec vectors [13]. Here, a serious drawback is the absence of an explanatory component – it's not possible to understand why the algorithm has made this or that particular decision.

In this paper, we propose our own solution to the profiling problem based on the study of the users' speech behavior in social media. We hypothesize that the psychological and social characteristics of a person are reflected in the vocabulary and semantics of the texts produced by that person. Based on this hypothesis, we propose a method that is similar in concept to the group of thematic modeling methods.

In order to reuse results of our previous work, we split the task into two closely related parts:

[2] Here, we understand a *genre* as a type of text that is being distinguished on the basis of an intention embedded in it. For example, gratitude, accusation, advice, request, etc.

[3] We rely on expert analysis here, but we're already on our way to develop automated semantic field extraction methods. However, this topic is outside the scope of the paper.

- Building a linguistic portrait of a user by identifying genres and semantic fields in their unlabeled texts
- Determining psychological parameters of the user based on the identified linguistic characteristics

This paper presents the methods used for the second part of the task.

In addition, we compared our method to the deep learning model proposed in [13] as some kind of baseline in order to test the quality of the proposed approach against it.

3. Proposed Solution

3.1. Dataset

The dataset consists of approximately 18,000 comments of 298 Russian social media users from the same SNS who were pre-surveyed to determine their psychological traits according to the Big Five model. All the users gave an informed consent for their data to be used in this research. The BFI characteristics were originally presented as floating point numbers according to the test scale, but for the purpose of our work and due to the relatively small sample size they were reduced to a binary (indicator) form using simple threshold cutoff. This also enables us to facilitate comparison with other studies, as they also tend to use binary form for the Big Five (BFI) traits levels.

The dataset structure is given in the Table 1.

"+" and "-" in the first column mean high and low BFI trait level respectively. Numbers represent the count of users with the specified level of the particular trait. The labels "bfia", "bfic", "bfio", "bfin" and "bfie" is a codification given for such traits as agreeableness, conscientiousness, openness to experience, neuroticism and extraversion and represents BFI personality traits in the exact same order.

	bfia	bfic	bfio	bfin	bfie
+	153	129	165	149	133
-	145	169	133	149	165

Table 1. Dataset structure

Text data (user comments) was preprocessed with the state-of-the-art natural language processing pipeline which included the TweeterTokenizer from the NLTK Python package [14] and our own dictionary-based normalizer trained on OpenCorpora.org dataset (https://opencorpora.org/).

3.2. Genres and Semantic Field Detection in User Texts

The detection of semantic fields and genres in user comments was performed by experts manually in the "Semograph" system. In total, 40 different genres and about 60 semantic fields (such as "question", "aggression", "jargon", "poetism", "common language", "foul language", etc) were identified in user texts.

3.3. BFI Detection based on Genres and Semantic Fields

At this step, we performed BFI detection for users based on their linguistic behavior.

We started with two matrices. The first one, which we call CF, represents the relation between all comments of all users and their corresponding semantic fields. This matrix was constructed from the data manually labeled by experts in the "Semograph" system. Each element CF_{fc} in it represents the number of occurrences of the specific semantic field f taken from all the fields F in the specific comment c taken from all the comments C. The second matrix, UB, represents each users' BFI levels calculated according to BFI test. It's elements UB_{tu} are real-valued numbers that show level of BFI trait t taken from the list of all traits T of the user u from the list of all users U.

First, we converted our real-valued matrix UB into a matrix of indicators I based on the mean value of each parameter:

$$I_{tu} = UB_{tu} > \overline{(UB_{tu})_{\forall u \in U}}, \forall u \in U, \forall t \in T \tag{1}$$

Thus, the matrix I element i_{tu} represents the indicator (0 or 1) for high level of BFI trait $t \in T$ for the user $u \in U$.

Next, user semantic profile S was constructed based on comments-fields matrix CF by averaging number of occurrences of semantic fields between comments of the same user:

$$S_{fu} = \overline{(CU_{fc})_{c \in C_u}}, \forall u \in U, \forall f \in F \tag{2}$$

where C_u denotes a list of the comments from the user u.

These values were normalized among the columns (per-user) to clamp the sum to 1, thus semantic profile matrix S can be viewed as a probability matrix representing the chance of encountering a specific semantic field $f \in F$ in the users' $u \in U$ comments.

At the last step, we combined S and I to get matrix B which effectively represents the probability of a highly accented BFI trait based on semantic fields' counts:

$$B_{tf} = \sum_{\forall u \in U} S_{fu} I_{tu}, \forall f \in F, \forall t \in T \tag{3}$$

This matrix B was then used to calculate BFI levels on the test dataset. The process for this follows the same math as above, but as a last step we multiply elements of users' semantic profile matrix S' by B to get an I', the matrix that show the users' BFI trait level:

$$I'_{tu} = \sum_{\forall f \in F} S'_{fu} B_{tf}, \forall u \in U, \forall t \in T \tag{4}$$

We should note here that while an original I matrix built on the train set is a binary indicator matrix, the resulting I' for the test set is real-valued and, in effect, a probability matrix. That makes our classifier fuzzy and allows us to make more flexible evaluation of the results.

Initial testing revealed the need to reduce the number of semantic fields used to determine the level of a specific BFI indicator, in particular to reduce computational costs. For this purpose, the selection of the set of semantic fields F, showing the best results, was carried out using a genetic algorithm. A minimum of the F1 metric among 3 experiments on a random test sample was used as a fitness function. The genes in the chro-

mosomes encoded the presence / absence of a specific semantic field. As a result, each separate psychological trait was matched with its own set of 25-35 language parameters that are relevant for the speech of each user.

Figure 1 displays the experiments' pipeline in the IDEF0 notation.

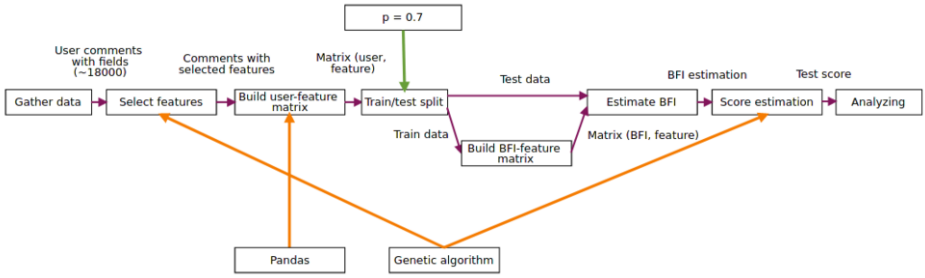

Figure 1. BFI detection pipeline.

The results of this stage of the experiment are shown in the Table 2.

3.4. BFI Detection with Deep Learning

An interesting model based on convolutional neural networks was proposed in [13]. That model solves the problem of user comment-based profiling using word2vec, MRC and LIWC. We decided to compare this model with the one we proposed. However, we were unable to directly use the code provided in our environment, so we implemented the model from scratch according to the described architecture using Keras [15], NumPy [16], Gensim [17], Theano [18] and TensorFlow [19] libraries. The LIWC Russian dictionary was also used.

Figure 2 shows a model of the experiment, also in the IDEF0 notation.

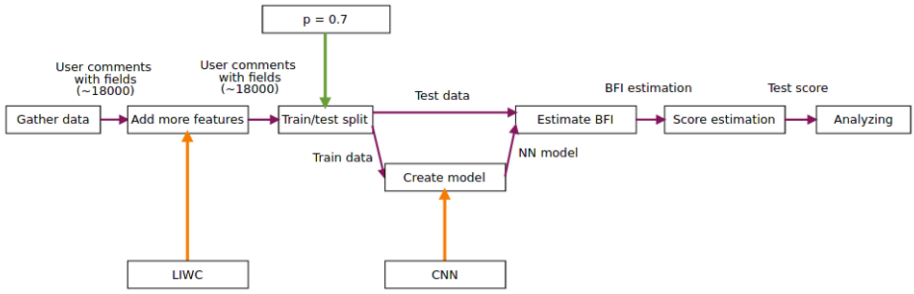

Figure 2. Experiment to identify BFI using a neural network

The results of this experiment are summarised in the Table 2.

3.5. Results

Table 2 shows the results of computational experiments as a F1-scores for each experiment (rows) and BFI traits (columns). CNN (E) – best original work results [13] for

	bfia	bfic	bfio	bfin	bfie
CNN (E)	0.56	0.57	**0.62**	0.59	0.58
CNN (R)	0.44	0.45	0.32	0.50	0.50
Our method	**0.69**	**0.65**	0.61	**0.70**	**0.59**

Table 2. Experiments' results, F1-score

English are taken as a baseline. CNN (R) – our implementation on our Russian dataset. Best results are highlighted.

We can see that in most cases our lingosemantic model scored even higher on Russian than original CNN on English, tailing close in the "bfio" case. Original CNN model performs quite low on the Russian language; perhaps some adjustments and tweaks may be of help here.

4. Conclusion

The detection of BFI parameters of social media users, as we have shown, can rely on linguistic analysis of speech behavior. Usage of a genetic algorithm for the selection of significant semantic fields for each psychological trait allowed to eliminate insignificant semantic fields for each psychological trait and significantly improved the profiling model with the BFI detection accuracy reaching 70%. Thus we can say that the presented approach to identifying the psychological parameters of social network users through a preliminary determination of their language behavior gives good results compared to the approaches proposed in the literature.

Overall, it can be concluded that the initial hypothesis about the influence of persons' psychological parameters on his linguistic behavior is proven to be correct, and the method can be used to efficiently build digital representations of SNS users, mark content and calculate optimal patterns of content movement in social media.

5. Future Work

The proposed method offers a wide scope for improvement. The next important step is to move from expert analysis in the data preparation procedure to an automated semantic field annotation methods. In the second stage of the algorithms, instead of using simple probability matrix, it is quite possible to use some more advanced algorithms. It is also possible to expand the list of the fields themselves, as well as apply boosting at the first stage to increase the volume of the training set by the comments of users who did not explicitly pass the psychological survey but gave an informed consent about participation.

Acknowledgments

This study is supported by the Ministry of Science and Higher Education of the Russian Federation, State Assignment No. FSNF-2020-0023 (Research Project of Perm State University, 2020–2022).

The experiment with the CNN was carried out on the computing cluster of the Perm State National Research University (https://wiki.hpc.psu.ru).

References

[1] Buffardi LE, Campbell WK. Narcissism and Social Networking Web Sites. Personality and Social Psychology Bulletin. 2008;34(10):1303–1314. PMID: 18599659. Available from: https://doi.org/10.1177/0146167208320061.

[2] Dunbar RIM, Arnaboldi V, Conti M, Passarella A. The structure of online social networks mirrors those in the offline world. Social Networks. 2015;43:39–47. Available from: https://www.sciencedirect.com/science/article/pii/S0378873315000313.

[3] Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, et al. What is a cognitive map? Organising knowledge for flexible behaviour. bioRxiv. 2018. Available from: https://www.biorxiv.org/content/early/2018/07/10/365593.

[4] Wilmer HH, Sherman LE, Chein JM. Smartphones and Cognition: A Review of Research Exploring the Links between Mobile Technology Habits and Cognitive Functioning. Frontiers in Psychology. 2017;8:605. Available from: https://www.frontiersin.org/article/10.3389/fpsyg.2017.00605.

[5] Baranov D, Belousov K, Erofeeva E, Leshchenko Y. SEMOGRAPH Information System as a Platform for Network-Based Linguistic Research: A Case Study of Verbal Behavior of Social Network Users. In: Uskov VL, Howlett RJ, Jain LC, editors. Smart Education and e-Learning 2019. Singapore: Springer Singapore; 2019. p. 313–324.

[6] Shchebetenko S. Reflexive characteristic adaptations explain sex differences in the Big Five: But not in neuroticism. Personality and Individual Differences. 2017;111:153–156. Available from: https://www.sciencedirect.com/science/article/pii/S0191886917300806.

[7] Belousov K, Erofeeva E, Baranov D, Zelyanskaya N, Shchebetenko S. The Multi-Parameter Analysis of Linguistic Data in the Information System Semograf (On the Example of the Study of Social Network Users' Speech). Vestnik Tomskogo gosudarstvennogo universiteta Filologiya. 2020 04:6–29.

[8] U VG, K S, Shenoy PD, R VK. An Overview on User Profiling in Online Social Networks. International Journal of Applied Information Systems. 2017 Jan;11(8):25–42. Available from: http://www.ijais.org/archives/volume11/number8/960-2017451639.

[9] Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology. 2010:0261927–09351676.

[10] Wilson M, Division I. MRC Psycholinguistic Database: Machine Usable Dictionary, Version 2.00. Behav Res Methods. 1997 06;20.

[11] Tandera T, H, Suhartono D, Wongso R, Lina Prasetio Y. Personality Prediction System from Facebook Users. Procedia Computer Science. 2017 12;116:604–611.

[12] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences. 2013;110(15):5802–5805. Available from: https://www.pnas.org/content/110/15/5802.

[13] Majumder N, Poria S, Gelbukh A, Cambria E. Deep Learning-Based Document Modeling for Personality Detection from Text. IEEE Intelligent Systems. 2017 Mar;32(2):74–79.

[14] Loper E, Bird S. NLTK: The Natural Language Toolkit. In: In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics; 2002. .

[15] Chollet F, et al.. Keras; 2015. https://keras.io.

[16] Oliphant T. NumPy: A guide to NumPy; 2006.

[17] Řehuřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010. p. 45–50. http://is.muni.cz/publication/884893/en.

[18] Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, et al. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints. 2016 May;abs/1605.02688. Available from: http://arxiv.org/abs/1605.02688.

[19] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Software available from tensorflow.org. Available from: https://www.tensorflow.org/.