Identification of user profiles in online social networks:
a combined approach with face recognition

V.D. Oliseenko¹ and M.V. Abramov²

¹ Junior Researcher, Assistant of the Department, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg State University, Saint-Petersburg, Russia
² Candidate of Technical Sciences, Head of Laboratory, Associate Professor, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg State University, Saint-Petersburg, Russia

Email: vdo@dscs.pro, mva@dscs.pro

Abstract. This paper suggests improving the previously existing method of identifying user profiles in different online social networks by adding face recognition results to the model. It is assumed that the method will become more stable for identifying people with the same name, city and age. It will help to find more user profiles in different online social networks, which will improve the estimation of their personal characteristics. Evaluating user personality traits is one of the key tasks in protecting employees of enterprises and companies from social engineering attacks.

1. Introduction
Unauthorized intrusions into information systems are becoming an increasingly acute problem for enterprises, companies and users of these systems. As noted by experts from Verizon [1], the number of attacks on information systems is growing at a huge pace from year to year. At the same time, the percentage of attacks using social engineering is also increasing. According to the study [2], social engineering was conducted in 84% of attacks on enterprises with the goal of installing malicious software. Considering the issue of the analysis of information systems' security from a software and hardware point of view, it can be noted that it is well studied and actively developed by various teams [3, 4, 5]. On the other hand, the question of analyzing the security and protection of users against social engineering attacks has been studied to a lesser extent. This is since the complexity of the study is much higher, as the object of study is a person, not a program or technique. This complexity is caused by the non-deterministic reactions of the user to various influences. In the same situation, a user may react differently to the same social engineering attack at different times. The user's response may depend on the user's personality, psychological state and other contexts.

To assess the extent of a user's personal characteristics, one can analyze the content published by the user in online social networks [6, 7]. In this case, the more such content can be extracted, the easier it is to build appropriate estimates. It is common to have user profiles in different online social networks. Aggregating more information from found user profiles in different online social networks can simplify the construction of estimates of personal characteristics of this user. Thus, the task of comparing user profiles in different online social networks in order to identify those belonging to the same person is relevant.
It should be noted that the task of comparing accounts in different online social networks is not new; there are approaches to its solution for different online social networks. For instance, [8] presents a systematic review of existing approaches and methods for defining profiles of one user in different online social networks. However, the presented approaches are applied to online social networks such as Facebook, Twitter, Foursquare and some others, which, although popular in Russia, are not among the top 3 most popular ones [9]. Many of the approaches presented in [8] are not applicable to the online social networks VK.com and OK.ru, which are the most popular in the Russian segment of the Internet [9]. For example, the profiles in the online social network Twitter do not have many attributes that are necessarily present in the online social networks VK.com or OK.ru, such as name, surname, city, age, etc. In addition to this, the task of comparing profiles in different social networks for different resources will be quite specific. At the same time, there is still no canonical method to solve this problem without errors and with a wide applicability.

The existing approaches oriented to the online social networks OK.ru and VK.com are based on comparing the values of user profile attributes to identify their belonging to one person [10, 11]. However, these approaches have some limitations that prevent them from achieving high enough accuracy (over 80%) and applicability. Thus, the task of automating the user profiles identification in different online social networks is relevant. One of the possible options to improve these approaches is to consider such important information from a user's profile as a photo. By finding faces on profile photos in different online social networks and adding the results of their comparison as a feature, for example, in the algorithm proposed in [11, 12], it is expected that it will be possible to improve the evaluation of identification and make it more stable.

The proposed article is devoted to the possibility of using facial recognition methods on users' photos in online social networks, in the context of using them in the approach [11, 12] of defining user profiles in online social networks VK.com and OK.ru.

2. Proposed Approach

According to the approach proposed in [11, 12], to determine the profiles belonging to one user, the problem of binary classification is solved, where $X$ is a set of mapped pairs of profiles of users of online social networks VK.com and OK.ru, and $Y$ is a set of classes $\{0; 1\}$, where $0$ means that a pair of profiles does not belong to one user, and $1$ means that it does. Since a binary classifier logistic regression is used, the features of this model are the numerical results of comparing the values of the corresponding attributes from user profiles, namely "surname", "name", "city", "age", "friends list". For comparing these attributes, the following metrics were selected:

- "surname", "city of residence" — Jaro-Winkler distance [13] $\delta_j = \frac{1}{3} \left( \frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{m} \right)$, where $m$ denotes matching word count, $s_1$ and $s_2$ denote compared string lengths and $t$ denotes the number of permutations;
- "name" — a dictionary with name forms; if there is no name in it, the Jaro-Winkler distance is used;
- "age" — is considered equal if the absolute difference is less than or equal to 3;
- "friends list" — user's friends are matched based on the "name" and "surname" attributes comparison by Jaro-Winkler distance with the cutoff threshold equal 0.8 and further the Shimkevich coefficient is calculated based on the found matching friends [12] $K = \frac{c}{\min(a, b)}$, where $a$ denotes the number of friends in the first profile, $b$ denotes the number of friends in the second profile, and $c$ denotes the number of matching friends from the first and the second profiles to determine the intersection of the friends lists in two profiles.

This work will consider the possibility of including face recognition results as a new feature to binary classification model. This inclusion is justified by the fact that it is expected to help to avoid false positives of the classification model when the owners of profiles are namesakes, live in the same city or have the same age. This may be relevant for people with common surnames and identical
names. For instance, only in the online social network VK.com there are 5339 people with the name and surname "Alexander Petrov", living in Moscow.

3. Existing Approaches
Recognition of faces in a photo consists of four stages [14]:
1. Find the faces in the photo;
2. Recognize every face, despite different lighting and picture angles;
3. Identify each special facial feature;
4. Compare faces of different people.

There are a lot of solution methods for each of the stages. However, we will consider the existing complex software solutions. For example, the article [15] gives an overview of popular computer vision libraries that allow solving each of the stages of face recognition. The authors of this paper consider the following libraries: OpenCV, OpenCV with IPP, LTI and VXL. After analyzing these libraries, the authors concluded that OpenCV is the best library for face recognition because it has the following advantages: speed of operation, flexibility in choosing the methods for each stage of face recognition, open source code, as well as cross-platform and cross-language implementation of the library. Using this library, it is planned to obtain the estimates of the probability that faces in different photos belong to the same person. Thus, the new attribute in the model [11] will be the probability that the faces from two avatars belong to one person, lying in the range [0,1].

4. Tools used
The work was done using the Python 3.9. The list of tools and libraries used in development is as follows:

- Python 3.9;
- Development Environment Jupyter Notebook;
- Library Pandas 1.1.14 and NumPy 1.19 for data analysis;
- Library Request to obtain data from online social networks;
- Library Matplotlib 3.3.2 for creating graphs.

5. Experiment Results
For the experiment, we used a data set from the previous work [11], consisting of 500 pairs of user profiles from the social networks VK.com and OK.ru, where 50% of pairs contain profiles belonging to the same user, and the other 50% to different users. We also used the logistic regression model which had been trained earlier. On this set, with the help of Python's face_recognition library [16], for each pair of user profiles the probability of belonging of faces from profile pictures to one person is calculated and the new model of logistic regression is trained. To test the result, 150 pairs of profiles belonging to one person and 50 pairs of profiles not belonging to one person were taken from the old data set, and other 100 pairs of profiles belonging to different people but having actual identical attributes "name", "surname", "city", and "age" were created. To estimate the quality of the classifier, the ROC curve [17] and its AUC area presented in Fig. 1, and the 4-fold cross-validation control procedure were used. The higher the AUC value is, the better the model is.
6. Conclusion
The new model with the result of comparison of faces showed a better result on the test data set compared to the old model, but for more reliable results, it is necessary to increase this data set. Further direction of work is to expand the data set and to implement this development into the existing set of programs to assess the security of users from social engineering attacks.

The paper presents an approach to the use of facial recognition methods on users' photos in online social networks, in the context of their use in the approach [11] for defining the profiles of the same users in online social networks VK.com and OK.ru. Aggregation of more information simplifies the construction of estimates of the user's personal characteristics, and indirectly helps to prevent social engineering attacks on users, companies and enterprises.

Acknowledgments
The research was carried out in the framework of the project on state assignment SPIIRAS № 0073-2019-0003, with the financial support of the RFBR (project №18-01-00626, №20-07-00839).

References
[1] Verizon Data Breach Investigations Report 2018. 2019. URL: https://www.researchgate.net/profile/Suzanne_Widup/publication/324455350_2018_Verizon_Data_Breach_Investigations_Report/links/5ace9f0b0f7e9b18965a5fe5/2018-Verizon-Data-Breach-InvestigationsReport.pdf?origin=publication_detail.

[2] Ptsecurity — Actual cyberthreats: results of 2019. 2020. URL: https://www.ptsecurity.com/ru/ru/research/analytics/cybersecurity-threatscape-2019/.

[3] Brantisky, A.A., Kotenko, I.V., Analysis and Classification of Methods for Network Attack Detection, SPIIRAS Proceedings (2016), 2(45), 207-244. doi: 10.15622/sp.45.13

[4] Kotenko, I., Parashchuk, I.? Verification of unreliable parameters of the malicious information detection model, Vestnik of Astrakhan State Technical University. Series: Management, computer science and informatics (2019), 2, 7-18. doi:10.24143/2072-9502-2019-2-7-18

[5] Zhang, Y., Zhang, L. Y., Zhou, J., Liu, L., Chen, F., A Review of Compressive Sensing in Information Security Field, IEEE Access (2016), vol. 4, 2507-2519.

[6] M.V. Abramov, T.V. Tulupyeva, A. L. Tulupyev, Social engineering attacks: social networks and user security estimates. SPb: PC GUAP, 2018, 266 p, ISBN 978-5-8088-1377-5.

[7] Kharitonov N.A., Maximov A.G., Tulupyev A.L., Algebraic Bayesian Networks: Naïve Frequentist Approach to Local Machine Learning Based on Imperfect Information from Social Media and Expert Estimates, Communications in Computer and Information Science, 2019.
[8] Hazimeh, H., Mugellini, E., Khaled, O. A., Cudré-Mauroux P. SocialMatching++: A Novel Approach for Interlinking User Profiles on Social Networks, In Proceedings of PROFILES’17@ ISWC. 2017.

[9] Statistics of social networks in Russia for 2018. URL: https://hiconversion.ru/blog/statistika-socialnyh-setej-v-rossii-na-2018-god/

[10] Russian Federation Patent № 2011145077/08, 08.11.2011. Way to integrate user profiles of online social networks

[11] Korepanova, A. A., Oliseenko, V. D., Abramov, M. V., & Tulupyev, A. L. (2019). Application of Machine Learning Methods in the Task of Identifying User Accounts in Two Social Networks. Computer Tools in Education, (3), 29-43. doi: 10.32603/2071-2340-2019-3-29-43.

[12] Korepanova A.A., Oliseenko V.D., Abramov M. V. Applicability of Similarity Coefficients in Social Circle Matching // 2020 XXIII International Conference on Soft Computing and Measurements (SCM). St. Petersburg, Russia. 2020. pp. 41-43. doi: 10.1109/SCM50615.2020.9198782.

[13] Winkler W. E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage // Proceedings of the Section on Survey Research Methods (American Statistical Association). 1990. P. 354–359.

[14] Al-Mukhtar F., Al-Dabagh M. Real-Time Face Recognition System Using KPCA, LBP and Support Vector Machine. International Journal of Advanced Engineering Research and Science. 2017. 4. 184-189. 10.22161/ijaers.4.2.36.

[15] Bulatnikov, E.V., Goeva, A.A. Comparison of computer vision libraries for application in an application that uses the technology of recognition of flat images, MSUE Newsletter, 2015, №6.

[16] Recognize and manipulate faces from Python URL: https://github.com/ageitgey/face_recognition

[17] Zweig, M.H., Campbell, G., Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, Clinical chemistry (1993), Vol. 39, № 4, 561-577.