Applications of artificial intelligence to identify fake accounts: psychological and legal aspects

Mikhail Smolenskiy[1,2] and *Nikolay* Levshin[1*]

[1] Don State Technical University, Gagarin sq., 1, Rostov-on-Don, 344003, Russia
[2] Rostov State Transport University, Rostovskogo Strelkovogo Polka Narodnogo Opolcheniya sq., 2, Rostov-on-Don, 344038, Russia

Abstract. This article analyzes the threats posed by the proliferation of fake accounts in today's information society. The authors consider the history of the emergence and development of this problem, as well as the ways proposed by governments and the international scientific community to address it. In search of a viable solution, they compare various methods for limiting the spread of fake accounts. They suggest using artificial intelligence technology as a promising alternative. As a training basis for AI, the authors propose analyzing the behavioral motivations of those who create fake accounts, in order to utilize a trained behavioral analytics system to combat fakes automatically. Along with key principles for using behavioral metrics, the authors provide various examples that illustrate the feasibility and effectiveness of their proposed solution. Special attention is paid to the legal aspects of implementing the proposed solution, as the process of detecting fake accounts is intricately linked to the processing of biometric and other personal data. The requirements for the storage and processing of personal data are becoming stricter every year. Therefore, in order to limit the spread of fake accounts, it is essential to find a solution that complies with information legislation.

1 Introduction

Stories of individuals who impersonate others or assume false identities in order to obtain benefits have been documented in human history since the earliest religious texts. Over the centuries, this theme has taken on many different forms. One of the most well-known examples is that of impostors who pretended to be relatives of rulers in order to seize power in medieval kingdoms. However, with the shift from a "man-to-man" interaction system to a "human-machine-human" one, the issue of deception has taken on a qualitatively different dimension and reached truly alarming proportions.

Anonymity and the use of fictitious information have provided numerous opportunities for internet users to engage in deviant behavior. Within modern information systems, an individual may maintain a significant number of virtual accounts under various names. These accounts may be designated as "fake", "virtual" or "sockpuppet", among other terms, which

* Corresponding author: info.law.expert@yandex.com

are often used interchangeably. In practice, the term "fake" is often preferred, as it is easier to pronounce and understand.

Fake accounts can be used for various purposes [1]. Below are some of the most common use cases for fake accounts:

1. Circumventing restrictions. By using fake accounts, individuals and organizations can circumvent blockages and sanctions imposed by various authorities. Additionally, fake accounts can be used to repeatedly receive discounts or free trial periods that are intended for one-time use, thereby circumventing distribution rules.

2. Manipulation of public opinion. Fake accounts can manipulate the results of polls and votes [2], as well as distribute advertising and propaganda [3]. They can also spread fake reviews that praise products, for example [4]. Furthermore, fake accounts may disseminate slander and misinformation, engage in political trolling [5], etc.

3. Evasion of legal responsibility. Some individuals use fake accounts to extort funds or distribute intellectual property without authorization [6]. Others may distribute prohibited information or unlicensed software.

4. Threat to life and health. A few examples of this include the distribution of narcotics [7], harassment, and violation of privacy. There may also be discussion of the incitement of suicide [8, 9].

It is evident that the diverse range of threats posed by the use of false accounts makes the issue of the proliferation of fake accounts a significant concern.

The COVID-19 pandemic and armed conflict in the territory of Ukraine and neighboring regions have resulted in many people losing their regular sources of income and finding themselves in a challenging financial situation. This has motivated them to seek new ways to generate revenue and explore the internet more actively in search of "quick" money. Consequently, the need for a solution to effectively detect fake accounts has become increasingly urgent every year.

It should be noted that the use of fake accounts can be prevented by preventing people from using internet resources under someone else's name. All the necessary technologies for this have long been available and are actively used in various aspects of our lives. However, such a restriction could be seen as an infringement on the freedom to access the internet, which in some countries is already recognized as a fundamental right. Moreover, implementing such a measure would require coordinated efforts and significant funding. Consequently, the prospect of complete personalization of information on the internet seems unlikely and unrealistic.

Gaps in current legal frameworks present a separate challenge. Individuals affected by the activities of fraudsters often do not even seek recourse through law enforcement agencies, as the cost of doing so may only serve to further exacerbate the harm they have experienced. Consequently, owners of online platforms must seek innovative solutions to deter criminal activity.

A new phase in the evolution of the issue of fake accounts emerged when attackers began employing artificial intelligence (AI) technologies to generate and maintain activity on accounts. Since then, the challenge of countering the proliferation of fake accounts has become akin to fighting a mythical Hydra, with each severed head being immediately replaced by several new ones. At this juncture, it has become apparent that an automated solution is necessary to identify fraudulent accounts, and the utilization of AI technologies appears to be the most promising avenue in this regard.

Researchers from around the world have proposed various solutions for both analyzing images within accounts and examining various metrics for account creation and operation. For instance, some researchers propose using deep machine learning to identify similar semantic patterns [10, 11]. However, this method appears to be ineffective due to the high

computational cost and high probability of error in identification. To enhance its effectiveness, each user would need to generate large volumes of text. Clearly, this solution may not be suitable for all services.

Our research began with the premise that by identifying behaviors indicative of users seeking to achieve destructive goals, it may be possible to use this information to identify the accounts of such users and proactively limit their capabilities before they have a chance to realize their goals.

It should be noted that the pattern of behavior in this instance and throughout the text refers to a relatively consistent sequence of behavioral actions (patterns) that describe the actions of an individual or group of individuals and are a response to specific types of situations (same for the individual in meaning, and, more specifically, in their personal context for similar situations) [12].

The behavior of a user of any information service is naturally constrained by the limits of interaction that are provided by the web interface designed by the developer. Consequently, the range of possible actions that a user can take is also limited. This enables the identification of characteristic patterns of behavior and the correlation of these patterns with the objectives of users. While this task may not be trivial, it is nevertheless feasible, even if each user possesses a unique and distinct personality.

Based on the above, the following research hypotheses have been formulated:

- the behavior patterns of users of electronic resources differ depending on the goals they seek to achieve.
- the patterns of behavior exhibited by intruders can be sufficiently distinct to allow them to be differentiated with a high degree of probability from those of other users.
- in the operation of multi-user information systems, it is possible to identify a set of metrics that, when analyzed by deep machine learning, will allow predictions to be made about user behavior.

2 Materials and methods

To test main hypotheses, general scientific methods of theoretical analysis were employed, among which a key role was played by the hypothetical-deductive approach, which involved correlating the hypotheses of this study with established hypotheses from the field of criminal profiling theory. This study was conducted on the basis of fundamental principles of behavioral psychology, and as such, some of the techniques employed are drawn from the realm of psychological research. Specifically, to identify behavioral patterns unique to the creators of fake accounts, a content analysis of surveys completed by site users during account registration was conducted. Statistical analysis was then used to determine the significance of these metrics. Assessing the legality of applying behavioral metrics analysis across different jurisdictions necessitated the use of formal legal and comparative legal methods. The method of social and legal experiment was used in order to protect against gaps in legislation.

3 Results

During the study, the following groups of indicators were identified that can be conditionally distinguished within the actions taken to create and populate accounts for users with ill-intentional goals, based on their orientation:

- the desire to conceal one's "true identity", i.e., actions aimed at preventing individuals interested in stopping a user's malicious activities from identifying their

true identity (through the use of proxies, email anonymization services, and receiving verification calls via third-party services), i.e., those actions that enable anonymity.

- the intention to impersonate a regular individual (via the use of first and last names in place of pseudonyms, the inclusion of photos, and completion of insignificant aspects of the account).
- the desire to attract the victim's attention. For example, on dating sites, the use of attractive photos, mostly of women, often with a hint of sexuality (revealing or form-fitting clothing), indicating sexual interests (to spark sexual attraction), and a wide age range of preferences for partners (to attract more users). For freelance platforms, the use of professional photos, detailed descriptions of experience, and low prices.

4 Discussions

Many on these objectives, it is possible to identify certain behavioral patterns that are common among different individuals and can be used to identify fake accounts. The human mind is rational and seeks to conserve resources. Therefore, when carrying out certain types of repetitive actions, such as creating images for new accounts, each individual subconsciously follows a certain pattern. This results in accounts created by a single individual becoming recognizable within a larger group. To illustrate this more clearly, consider the data presented in Table 1.

Table 1. List of registered account IDs.

№	ID
1	marussiasem
2	sandimqa
3	weetaaa
4	k_s_e_n_i_a__036
5	id_a_r_c_h_i_e
6	khanyuriy851
7	aaermolova

At a glance at Table 1, one can observe the general pattern in the creation of identifiers 4 and 5. While it is possible for different individuals to use the underscore character as a separator in an identifier, it is unlikely that two individuals would independently choose this pattern when creating accounts within a short timeframe in the same database. It is more likely that these accounts were generated by the same individual. Further analysis of the patterns in the profile information of these accounts may help confirm or refute this assumption, depending on whether there are other coincidences that support or contradict the hypothesis of a single individual creating both accounts.

Using artificial intelligence to populate profiles can be detrimental to fake account creators. This is because, as a result of generating profiles based on a specific template, a large number of potential matches may be generated [13]. To support this assertion, let us consider screenshots of user profile data from the database of a dating service (see Figures 1 and 2).

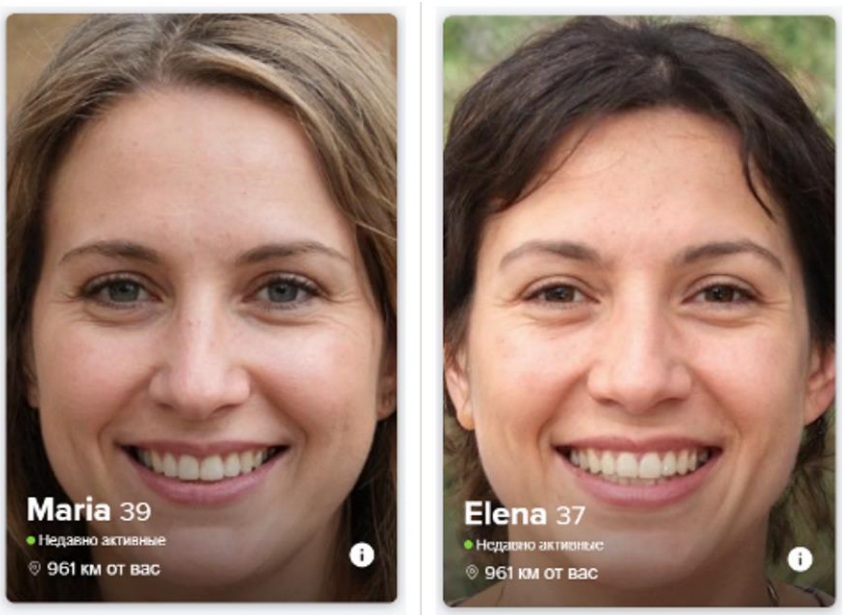

Fig. 1. Comparison of profiles of light-skinned fake profiles on a dating service

Figures 1 and 2 demonstrate the similarities in the methods of completing different forms. Firstly, the relevant age range and choice of photographs are identical. The questionnaires in question use portrait photographs that are similar in style to an identikit.

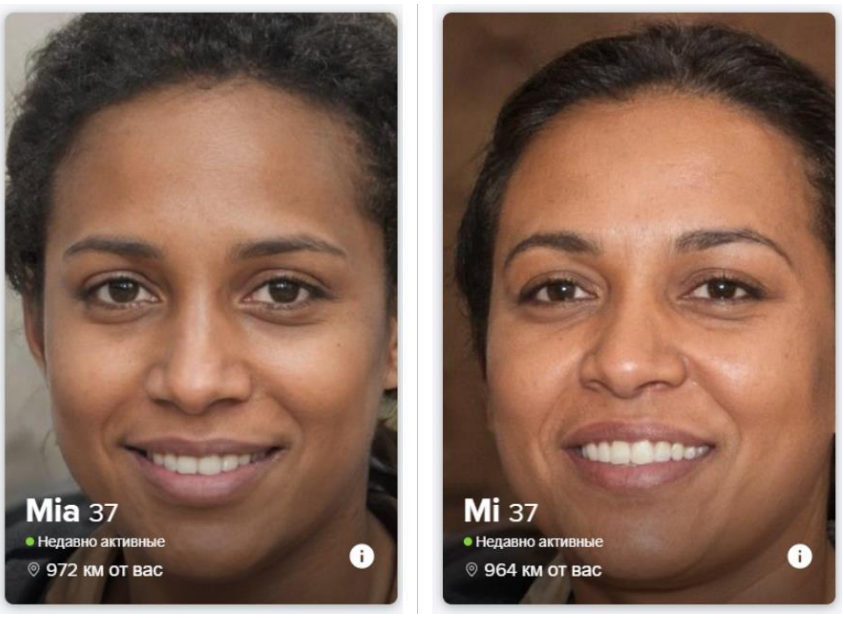

Fig. 2. Comparison of profiles of dark-skinned fake profiles on a dating service

The eyes, nose, mouth, and chin are aligned on a single line, which is clearly visible when comparing them in pairs. This alignment is due to the neural network that generates the images, which collects them as sketches of elements of similar size.

The creators of fake accounts often mimic the practices of advertising image producers and PR professionals, attempting to select an image that would endear them to their intended audience. Therefore, they endeavor to make their profiles as realistic as possible. It is not necessary for a fake account to conceal a photo or utilize a pseudonym, as they do not value fictitious data. On the contrary, the creator of a fake profile is interested in their potential victim perceiving them as a person who exists in the real world. Consequently, the creator of the deceptive profile endeavors to include as many details as feasible in their public profile.

At the same time, attackers are attempting to conceal their true identity from the administrators of the online platform as much as possible. They do this by using fake email addresses and proxy servers to create accounts. In contrast, ordinary users typically use their real information to register accounts, although in public profiles, they attempt to provide as little identifying information as possible to other users.

By building enhanced behavioral models, it is possible to identify user actions that are most indicative of the pursuit of a specific objective. For instance, for a spammer, this objective might be posting a link, while for a "puppeteer" using fabricated identities in order to manipulate the results of a survey, participating in voting.

It is of particular interest to analyze patterns in the behavior of individuals who engage in fraudulent activities, such as extorting money or other benefits through various means. These individuals are interested in deceiving potential victims into believing in the persona they create. Therefore, when creating a fake user profile, they attempt to present a likable and trustworthy image that would be well-received by their intended audience.

Understanding this principle of fraudsters' operation allows one to significantly narrow down the "circle of potential suspects". For instance, in the context of online dating services, it is important to recognize that due to societal stereotypes, it may be easier for a scammer to lure funds by posing as a woman rather than a man, given the existing gender-based biases. Therefore, it is advisable to focus the search for scammer accounts primarily on women's profiles [14].

To attract the attention of potential victims, the scammer may attempt to include explicit photographs or text in the questionnaire that a person concerned about their reputation would not typically include. In order to make the fabricated profile appear more credible, the fraudster may include a last name in the questionnaire as they are not concerned about the privacy or moral integrity of the individual they have created.

A decade ago, the automation of the detection of such "blurred" features seemed like an impossible task. However, the advancement of artificial intelligence (AI) technologies in the fields of text analysis and visual image recognition has allowed us to re-examine this challenge and find an effective solution to the issue of preventive detection and neutralization of network threats.

It is essential to consider the potential errors of the models. For instance, an account may have been voluntarily transferred or sold, or unauthorized access may have been gained to it. Clearly, in these instances, most of the metrics would not align with those typically associated with intruders. Due to the fact that, up until this point, the account holder has acted as a legitimate user, there will be no indications of an attempt to conceal their identity within the account. Consequently, the system of behavioral metrics could face difficulties in identifying intruders in such circumstances.

In the past, the use of proxy servers and disposable email addresses was a sign that a user had malicious intentions. However, today, they have become more common among ordinary people. Virtual private network (VPN) services are used by individuals who need to access certain resources in order to circumvent restrictions. Disposable email addresses can be used by those who wish to protect themselves from data breaches or prevent their contacts from being added to spam databases. As a result, these technical indicators can no longer be considered reliable in the same manner as they were in the past.

It should also be noted that errors specific to individual services may occur. For instance, on dating platforms, users may encounter profiles of individuals who offer sexual services in exchange for monetary compensation. Similar to the creators of false profiles, these individuals seek to conceal their true identities and obtain advantages. Consequently, the content of their profiles will align with the conduct of the creators of fake accounts in several metrics. Nevertheless, formally, such profiles will not be classified as fraudulent. However, on this case, such an occurrence does not represent a significant issue, as prostitution is prohibited in the majority of jurisdictions and the operators of dating platforms are equally motivated to block such profiles as they are to eliminate fake accounts.

Along with potential errors, there is another significant issue that needs to be addressed when utilizing behavioral analysis of user activity. This issue relates to compliance with regulatory standards for the storage and processing of personal information [15, 16]. The regulations governing personal data are becoming stricter each year. Additionally, each region may have its own unique set of regulations regarding personal data, which can create challenges for both research and information service providers. Therefore, in order to limit the proliferation of fraudulent accounts, it is essential to find a solution that abides by the requirements of data protection legislation.

It is important to consider that the laws of each jurisdiction have their own definition of personal information. These definitions may vary slightly, but in most cases they are similar in that they define personal information as data that can be used to identify an individual.

As part of efforts to combat the proliferation of fake accounts, the aforementioned situation presents a potential challenge. This is due to the necessity of addressing the tasks of constructing a psychological profile of the owners of such accounts and identifying markers that permit the rapid identification of the accounts they generate. In essence, we are seeking to identify individuals, and as such, we operate with personal information. At present, the only means to ensure the legitimacy of this process is through data encryption [17, 18].

5 Conclusion

To summarize, based on the findings of the research, a number of behavioral indicators have been identified that can be used to detect fake accounts with a high level of accuracy even during the account registration process. However, it is not practical to publicly disclose these indicators, as attackers could use the information to improve their methods for creating and registering fake accounts. Instead, the solutions developed as part of the research will be made available in the form of a pre-trained neural network model that can be integrated into multi-user information systems.

References

1. N. Alieva et al., Lecture Notes in Networks and Systems, **57**, 1020-1026 (2019) doi: 10.1007/978-3-030-00102-5_108
2. M. Orabi et al., Information Processing & Management, **57**, 102250 (2020) doi:10.1016/j.ipm.2020.102250
3. O. Varol, Online Social Networks and Media, **37–38**, 100263 (2023) doi:10.1016/j.osnem.2023.100263
4. C. Esposito et al., Computers & Security, **133**, 103407 (2023) doi:10.1016/j.cose.2023.103407
5. V. Zvereva, Russian Literature, **118**, 107-140 (2020) doi:10.1016/j.ruslit.2020.11.005

6. J. Cabrero-Holgueras, S. Pastrana, Computers & Security, **111**, 102489 (2021) doi:10.1016/j.cose.2021.102489
7. O. Grechenkova, Y. Kuzmenko, Lecture Notes in Networks and Systems, **57**, 612-621 (2019), doi:10.1007/978-3-030-00102-5_64
8. Y. Isakova, E. Millerov, E3S Web of Conferences, **273**, 08083 (2021) doi:10.1051/e3sconf/202127308083
9. J. Uyheng et al., Information Processing & Management, **59**, 103012 (2022) doi:10.1016/j.ipm.2022.103012
10. X. Du et al., Expert Systems with Applications, **207**, 117924 (2022) doi:10.1016/j.eswa.2022.117924
11. Z. Yamak et al., Knowledge-Based Systems, **149**, 124-142 (2018) doi:10.1016/j.knosys.2018.03.002
12. I. Abakumova et al., E3S Web of Conferences, **210**, 20015 (2020) doi:10.1051/e3sconf/202021020015
13. J. Mehta et al., Expert Systems with Applications, **219**, 119669 (2023) doi:10.1016/j.eswa.2023.119669
14. P. Anesa, Discourse, Context & Media, **35**, 100398 (2020) doi:10.1016/j.dcm.2020.100398
15. J. Meszaros, Ch. Ho, Computer Law & Security Review, **41**, 105532 (2021) doi:10.1016/j.clsr.2021.105532
16. M. Smolenskiy, N. Levshin, E3S Web of Conferences, **273**, 08099 (2021) doi:10.1051/e3sconf/202127308099
17. V. Terziyan et al., Procedia Computer Science, **232**, 2201-2212 (2024) doi:10.1016/j.procs.2024.02.039
18. V. Galushka et al., Journal of Physics: Conference Series, **1015**, 042003 (2018) doi:10.1088/1742-6596/1015/4/042003