Machine Learning-Based malicious users detection in the VKontakte social network

Denis Samokhvalov
National Research University Higher School of Economics
disamokhvalov@edu.hse.ru

***Abstract*—This paper presents a machine learning-based approach for detection of malicious users in the largest Russian online social network VKontakte. An exploratory data analysis was conducted to determine the insights and anomalies in a dataset consisted of 42394 malicious and 241035 genuine accounts. Furthermore, a tool for automated collection of the information about malicious accounts in the VKontakte online social network was developed and used for the dataset collection, described in this research. A baseline feature engineering was conducted and the CatBoost classifier was used to build a classification model. The results showed that this model can identify malicious users with an overall 0.91 AUC-score validated with 4-folds cross-validation approach.**

Keywords*—*VKontakte, malicious users, machine learning, social networks, classification models.

I. Introduction

An online social network (OSN) is an online platform that allows people who share the same views or have real-life connections to interact with each other online [1]. OSNs also provide users with a great ability to communicate, entertain, consume and share a different type of information that they are interested in. Moreover, modern social networks have become the platforms where companies can promote and even sell their products while maintaining good relationships with their customers through clear communication channels [2, 3].

Being a great instrument for connecting people and hosting useful information, OSNs try to attract as many users as possible, thus a strong authentication (by personal ID or driving license for ex.) is not required for an account creation as a rule. For example, in the OSN VKontakte, for a long time, it was possible to register an account by submitting only an e-mail address. VKontakte team made an authentication by mobile telephone number required for a valid account creation, however, this still does not fully solve the issue, since it is possible for a malicious identity to use multiple sim-cards or so-called virtual numbers [4].

Lack of strong authentication provides an opportunity for malicious users to evade OSNs with malicious activity, such as spamming, phishing, distribution of malicious software, trolling, terrorism and others [5–8]. While these are the activities that evaded the internet almost since its invention, several new threats relevant to OSNs have appeared [9, 10]:

- *Clickjacking* - a malicious practice where a user is made to click on something that behaves not the same way as it should to the prior knowledge of the user.
- *Crowdturfing* - a campaign that aims to gain or destroy the reputation of people, products and other entities through spreading biased opinions and framed information.
- *Fake account attack* - a most commonly used type of attack when an account with fake credentials created for interaction with the legitimate users.
- *Identity clone attack* - a malicious practice where an attacker creates a new fake profile while using stolen private information of an existent user.
- *Cyberstalking* - harassment of an individual in the social network.

The aforementioned threats are relevant for most of the existing social networks and in most cases, they are performed by fakes.

Facebook, the largest social network in the world, reports that 8.7% of its accounts which amounts to approximately 206 million do not belong to real users [11]. For addressing this vital issue Facebook even created its security system for protecting users from malicious activity and it is known as Facebook Immune System (FIS) [12]. While being a scalable real-time system that can process hundreds of thousands read and write actions per second, it cannot still detect all the types of malicious activity [13, 14].

The goal of this research is to analyze the application of machine learning techniques for the detection of malicious users in OSN VKontakte. The information about the total number of 42394 malicious accounts was collected with the help of developed automated VK-scraper tool. In this research, we show that VKontakte malicious users have a specificity that is possible to use for building a highly accurate classification model.

The main contributions of this paper are the following:

- We propose an architecture for automated malicious accounts collection tool called VK-scraper.
- An exploratory data analysis of malicious VKontakte accounts was conducted and the main differences between malicious and genuine accounts were revealed.
- We show that Catboost performs better than Neural Nets approach proposed by other researcher for this problem.
- We provide a benchmark of the most important features identified by Catboost.

The outputs of this paper can be used further by other researches of malicious activity in VKontakte OSN.

II. Related Work

The machine learning-based detection of malicious users in OSNs has attracted the attention of both researches and businesses when machine learning became an industrially popular and valuable approach. In [15] an application of Matrix factorization and SVM for spam accounts detection in Chinese OSN Renren was proposed. In this work, authors collected a dataset out of 33116 accounts, manually classified them into spammers and non-spammers and applied the SVM algorithm for spammers detection on a set of messages content and users' social behavior. They managed to reach an outstanding performance with a true positive rate of spammers detection reaching 99.1%. The Longitudinal Data Analysis of the Social graph method for the detection of so-called Friends farms in VKontakte was developed in [16]. This work aimed to detect fake identities among newly registered users of vk.com. According to conducted longitudinal analysis, authors revealed that fake profiles are more likely to be found among those users that show abnormal behavior in the growth of social graph metrics such as degree, reciprocated ties and clustering. In [14] a framework for detecting Fake account attacks on Facebook was described. The research studied the temporal evolution of OSNs and the characteristics of the real users' profiles. Researchers presented a way to analyze social network graphs from a dynamic point of view within the context of privacy threats. The application of machine learning techniques for fake profiles identification in Linkedin was described in [17]. Since Linkedin is a quite closed OSN that does not expose any API to the outer world, it is rather hard to get any data for the analysis from there. Authors of this work showed that even having a very limited dataset of only 27 fake accounts, it is possible to achieve a result comparable to the results obtained by other existing approaches based on the larger data set. An instrument called SybilRank was developed in [18]. SybilRank is used for detecting the fake users (called Sybils) in Tuenti OSN by analyzing the social graph properties. The developed tool allowed to achieve at least 20% lower false positive and negative rates than the second-best contender in most of the attack scenarios. Sophisticated techniques for data normalization and noise removals such as Artificial Bee Colony (ABC) and Ant Colony Optimization (ACO) were used in [19] among which 3 supervised machine learning algorithms (Naive Bayes, SVM, and Decision Trees) were applied to predict the fake users' profiles on Facebook. The CRAWLER tool was developed in [20] and a total number of 992 profiles were crawled with the help of this tool, out of which 201 turned out to be malicious. An application of both supervised (Decision Trees, KNN, SVM) and unsupervised (K-means, K-medoids) machine learning algorithms were used for classification, and a decent qualities of the models were obtained. In [21] an application of methods such as PCA, Spearman's Rank-Order Correlation, Wrapper Feature Selection using SVM is described for dimensionality reduction to reduce the number of low-importance features for the fake accounts' detection in the social media. In the research, several existing datasets of both real and fake Twitter accounts, crawled by other researchers, is used. A set of feature selection techniques was evaluated to achieve the best performance and classification results. In [22] an analysis of the tonality of the statuses of users of the OSN Facebook is conducted. Authors compared machine learning algorithms Naïve Bayes, Rocchio, and multi-layer perceptron by applying them on the 7000 status updates received from 900 Facebook users. All of the statuses were manually divided into two classes: positive and negative, however since there were significantly fewer negative reviews in the sample, the authors used 1131 reviews of each class to balance the classes in the final training dataset. In [12] a software application and architecture described. The application aims to protect users and the social graph from malicious actions by cybercriminals. The described system operates in real-time and, according to the statements of its creators, checks and classifies each read and write action. As of March 2011, the system performed 25 billion checks per day, with a peak frequency of 650,000 checks per second. Authors of [23] describe an approach to identify automatically managed accounts or so called bots in the VKontakte OSN. Authors use a feedforward neural network and a sample of 4918 blocked accounts to train the model that shows a decent result on the validation set. Authors use an approach for sampling malicious accounts that is similar to one described in this paper, however the method they use in their research is not automated and thus can not be done in a standalone way. There is now evidence of what features turned out to be the most important and also it is not clear how exactly status-based features were generated. In [24] authors explore stacking ensemble approach on top of a combination of different types of models that were trained on the attributes of three different types: friendship graph, subscription information and user's texts. The result received in this article is 4-9% better than in [23]. In [25] a framework for extracting a large collections of Twitter accounts was proposed. Based on these features, several highly accurate models were built and their performances were evaluated on both an existing public dataset and an additional sample of manually-annotated Twitter accounts collected with a different strategy. Based on the models predictions, authors evaluated that percentage of Twitter accounts exhibiting social bot behaviors is between 9% and 15% and the behaviour of such accounts can be detected by supervised machine learning techniques. In [26] a model which increases the recall in detecting bots, allowing a researcher to delete more bots in Twitter, was proposed. Authors proposed an algorithm called Boosting through Optimizing Recall which was applied on top of a combination of twitter-specific heuristic features and features obtained through topic modelling of the tweets. The algorithm showed a result relatively better than other state-of-the-art models like AdaBoost.

III. Proposed Method

In this paper, a description of state-of-the-art machine learning techniques application for malicious users' identification in VKontakte OSN is presented. Moreover, an automated tool VK-scraper for scraping the data about malicious accounts before their actual removal by VKontakte administration is developed and its architecture is described in this research. A sample of 42395 of actual malicious users was collected and a set of data and feature engineering techniques were applied before the actual ML-model training.

A. VK-scraper

One of the most challenging parts of the malicious accounts detection domain is data collection. Even though some OSNs provide a useful API for the developers to interact with

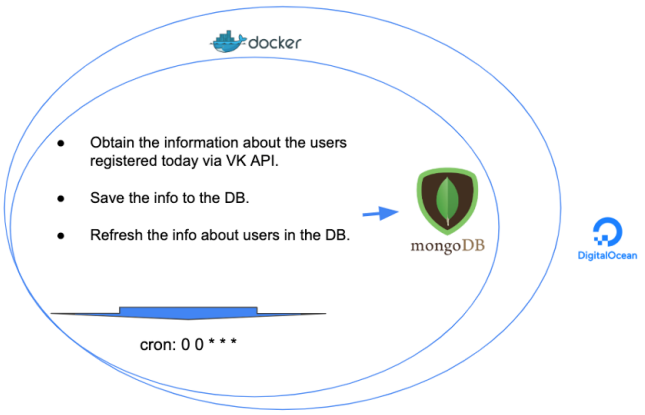

Fig. 1. VK-Scraper architecture

the platform and query the publicly available data, there is still a lack of techniques that allow gathering the available information about the blocked accounts since this data is not exposed by OSNs to the outer world after an account was blocked for a malicious activity. There were some workarounds proposed by researchers to deal with this obstacle, for example, expert evaluation, manual labeling, friends connections crawling, social graph properties analyses, etc. [16, 17, 20]. As was noted in [17] VKontakte assigns a unique incremental id to every user that is registered on the platform, thus it is easy to reverse engineer the relative timeline of VKontakte accounts registration. Since most of the malicious accounts are manually banned by VKontakte administration (due to the legitimate users complains mostly) within the first week of their existence, it is quite hard to detect a malicious user among the users that were registered a long time ago.

VK-scraper tool works in the following way. Every day it checks if there were any changes in the data that are stored in the VK-scraper MongoDB [27] database by simply calling the VKontakte API and comparing the data from the response to the data stored in the database. If there was a change, for example, a user updated its status or has been banned by the administration, it updates the information in the database by changing the differing fields. After that, it collects the information about 120,000 newly registered accounts in VKontakte by simply iterating over the 120,000 largest accounts ids that exist in the OSN. The newly scraped ids are stored in the VK-scraper database.

MongoDB was used as a local DB for storing data as it perfectly suits for storing JSON data and does not require a schema.

VK-scraper is wrapped with Docker [28] and deployed on a dedicated VPS provided by DigitalOcean developer cloud [29].

VK-scraper worked for 30 days (from 01.10.2019 to 30.10.2019) on a dedicated VPS and collected information about 3.5 million accounts, out of which 42394 turned out to be malicious.

B. Feature Engineering

VKontakte API provides access to query all the publicly available information about any open VK account. For example, it is possible to get information about the schools or universities that a specific user attended or what types of music she prefers if this data is provided by the user. Most of the accounts features available via VK API are categorical. The categorical feature is a feature that has a discrete set of values that are not necessarily comparable with each other (e.g., user ID or name of a city) [30]. Unfortunately, the number of values that are relevant for some feature can be quite large (for example, there are more than 200 countries available for selection during registration in vk.com) and this can make the model training and evaluation quite hard and even biased if the training dataset is limited and cannot cover all the available values. Thus, unlike other approaches specified in [23, 24], it was decided to convert all the categorical features into binary which are simply the indicators of whether this feature was specified by the account holder or not.

C. Catboost

There are plenty of machine learning algorithms for solving a binary classification model available today. One of the most robust is gradient boosting algorithm [30]. Catboost [30] – is an open-source library developed by Russian tech-giant Yandex that implements gradient boosting algorithm with special orientation on performance and processing of categorical features. It outperforms other popular implementation of gradient boosting in terms of quality on the classification tasks.

IV. Experiments and Results

Before building the actual model, an Exploratory Data analysis was conducted to compare the malicious and genuine user datasets and find the anomalies or extract the insights from the data. After that, a CatBoost model was trained on 4-fold cross-validation with the Log-Loss metric optimized on the fly.

A. Exploratory Data Analysis

A comparison of malicious and genuine accounts dataset revealed that there is a larger portion of genuine users who has certain info fulfilled in their profiles rather than malicious users (Fig. 2). For example, 57% of genuine users specified the country they currently live in their profiles, compared to 28% for malicious accounts; 40% of genuine users indicated the schools they studied in, while only 15% of malicious accounts had this information in their profiles. Most of the malicious accounts (78%) have female sex and also most of them (81%) have at least one photo uploaded. 36% of malicious accounts has at least one friend. Two most popular professions are entrepreneur and princess. It was also revealed that 98% of malicious users have their mobile phones connected to their accounts, while only 59% of genuine users linked their phone numbers to their profiles. After researching for a while about that, it was found out that until a certain time, it was possible to register in VKontakte without having a phone number linked to the account during the registration process, however, nowadays it is impossible to create an account without having a mobile phone number assigned to the actual account. Since the

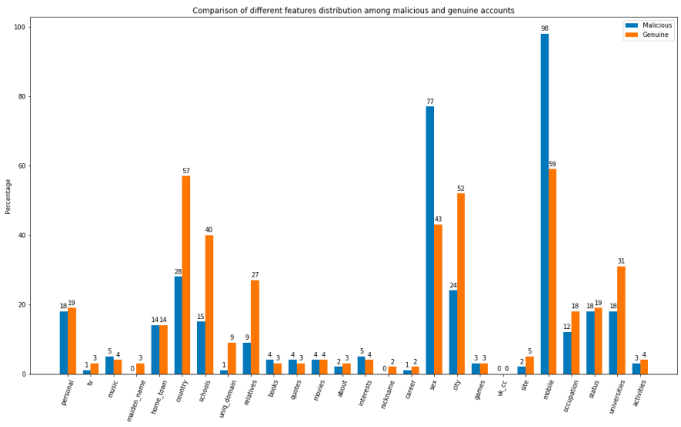

Fig. 2. Comparison of different features distribution among malicious and genuine accounts

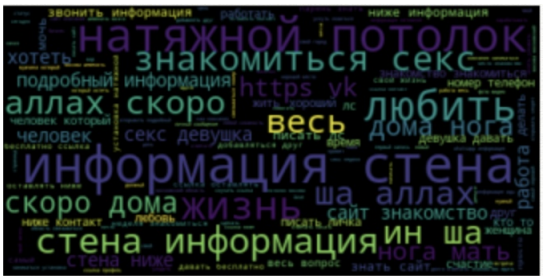

Fig. 3. Malicious users statuses wordcloud. The larger the word the more often it occurs in the statuses of malicious accounts.

malicious users' dataset contains only newly registered users and the genuine dataset is by one half a random selection from all of the existing accounts, it was decided to remove this feature from the both of the datasets when training the model to prevent overfitting on the peculiar properties of our data.

Another interesting part that requires a more detailed exploration is user statuses. Status - is a short (less than 140 symbols) text that a user can outline right below his profile name on the main page. It was found out that roughly 19% of malicious accounts have a status specified on their page and the meaning of those statuses is generally malicious. The biggest portion of the accounts promote sexual services or contain extremist slogans. It might be interesting to research on NLP techniques application for detection of malicious statuses in the future.

B. Model training

The CatBoost classifier was trained on the given dataset with 4-fold cross-validation, with 20% of the dataset belonging to a fold on each training round (Fig. 4).

As can be seen, a baseline of 0.91 was achieved on 4-fold cross-validation which is 15 points better than in [23]. A benchmark of feature importances is provided in Table II.

V. Further Directions

Further research concerns both technical and performance improvement of the VK-scraper tool and investigation of more advanced supervised learning techniques such as Deep Learning and NLP. It also might be interesting to try to apply some rotating proxy services such as Micro-leaves [31] for bypassing VKontakte API query restrictions and a larger-scale network scanning. This approach might require a larger-scale computational resource.

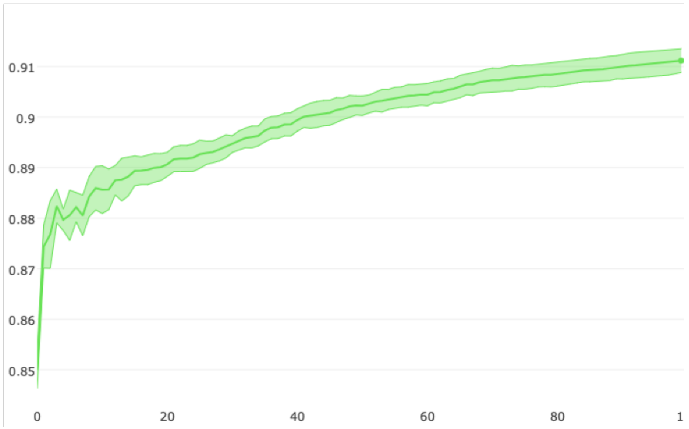

Fig. 4. AUC learning curve

	f0l	f1l	f2l	f3l	f0t	f1t	f2t	f3t
curr.	0.89	0.89	0.89	0.89	0.88	0.89	0.89	0.89
best.	-	-	-	-	0.90	0.91	0.91	0.91

TABLE I. 4-fold cross-validation AUC scores

The application of Deep Learning models can be considered after a larger sample of malicious users will be gathered. For example, it could be interesting to predict the probability of an account being malicious by applying Convolutional Neural Networks to profile photos of the gathered accounts. Another interesting direction is the application of NLP techniques for analyzing the accounts statuses and wall posts. It might be not such a good standalone solution since the majority of accounts do not have any status in their profiles and the daily quota for the wall API requests is limited to 5000 per day, but within an ensemble of Machine Learning and Deep Learning models, it could make a valuable contribution to the overall ensemble score.

VI. Conclusion

In this work, a Machine-Learning based approach for detection of the malicious users in the VKontakte OSN was presented. 42394 malicious users were scraped with the developed automated tool called VK-scraper. An exploratory data analysis for both malicious and genuine datasets was

Feature	Importance
bdate-day	22.335
sex	15.46
has-country	12.638
has-schools	8.989
has-relatives	8.532
friends-number	7.648
unique-domain	2.738
has-city	1.674
has-maiden-name	1.026
has-status	0.417

TABLE II. Feature importances

conducted and revealed that there is an evident difference between malicious and genuine VKontakte accounts. While the result of 0.91 AUC-score looks promising, there is still a room for improvement where more sophisticated techniques such as Deep Learning and NLP might come in.

REFERENCES

[1] J. A. Obar and S. S. Wildman, "Social Media Definition and the Governance Challenge: An Introduction to the Special Issue," *SSRN Journal*, 2015. [Online]. Available: http://www.ssrn.com/abstract=2637879

[2] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Proceedings of the 20th international conference companion on World wide web - WWW '11*. Hyderabad, India: ACM Press, 2011, p. 113. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1963192.1963250

[3] J. A. Obar and S. Wildman, "Social media definition and the governance challenge: An introduction to the special issue," *Telecommunications Policy*, vol. 39, no. 9, pp. 745–750, Oct. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0308596115001172

[4] I. Shatilin, "What are virtual SIM cards and what do they do?" [Online]. Available: https://www.kaspersky.com/blog/virtual-sim/11572/

[5] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: Dark of the social networks," *Journal of Network and Computer Applications*, vol. 79, pp. 41–67, Feb. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1084804516303009

[6] A. V. Filimonov, A. V. Osipov, and A. B. Klimov, "Application of neural networks to identify trolls in social networks," *arXiv:1504.07416 [cs]*, Apr. 2015, arXiv: 1504.07416. [Online]. Available: http://arxiv.org/abs/1504.07416

[7] A. Malm, R. Nash, and R. Moghadam, "Social Network Analysis and Terrorism," in *The Handbook of the Criminology of Terrorism*, G. LaFree and J. D. Freilich, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2017, pp. 221–231. [Online]. Available: http://doi.wiley.com/10.1002/9781118923986.ch14

[8] Z. Mao, D. Li, Y. Yang, X. Fu, and W. Yang, "Chinese DMOs' engagement on global social media: examining post-related factors," *Asia Pacific Journal of Tourism Research*, vol. 25, no. 3, pp. 274–285, Mar. 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10941665.2019.1708759

[9] D. DeBarr and H. Wechsler, "Using Social Network Analysis for Spam Detection," in *Advances in Social Computing*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S.-K. Chai, J. J. Salerno, and P. L. Mabry, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, vol. 6007, pp. 62–69, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-12079-4_10

[10] L. Wu and H. Liu, "Detecting Crowdturfing in Social Media," in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj and J. Rokne, Eds. New York, NY: Springer New York, 2017, pp. 1–9. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-7163-9_110196-1

[11] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, "Friend or foe? Fake profile identification in online social networks," *Soc. Netw. Anal. Min.*, vol. 4, no. 1, p. 194, Dec. 2014. [Online]. Available: http://link.springer.com/10.1007/s13278-014-0194-4

[12] T. Stein, E. Chen, and K. Mangla, "Facebook immune system," in *Proceedings of the 4th Workshop on Social Network Systems - SNS '11*. Salzburg, Austria: ACM Press, 2011, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1989656.1989664

[13] S. Ali, N. Islam, A. Rauf, I. Din, M. Guizani, and J. Rodrigues, "Privacy and Security Issues in Online Social Networks," *Future Internet*, vol. 10, no. 12, p. 114, Nov. 2018. [Online]. Available: http://www.mdpi.com/1999-5903/10/12/114

[14] M. Conti, R. Poovendran, and M. Secchiero, "FakeBook: Detecting Fake Profiles in On-Line Social Networks," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Istanbul: IEEE, Aug. 2012, pp. 1071–1078. [Online]. Available: http://ieeexplore.ieee.org/document/6425616/

[15] II M.E., (CSE),, M. A. Banu, M. N. Ahamed, M.E., Assistant Professor,, M. Manivannan, M.E., Associate Professor,, M. Vanitha, M.E., (Ph.D.,) Assistant Professor, D. Musthafa, and M.Tech., Ph.D., Associate Professor Al-Ameen Engineering College, Erode, Tamilnadu, India, "Detecting Spammers on Social Networks," *IJECS*, Feb. 2017. [Online]. Available: http://ijecs.in/issue/v6-i2/14%20ijecs.pdf

[16] A. Romanov, A. Semenov, and J. Veijalainen, "Revealing Fake Profiles in Social Networks by Longitudinal Data Analysis:," in *Proceedings of the 13th International Conference on Web Information Systems and Technologies*. Porto, Portugal: SCITEPRESS - Science and Technology Publications, 2017, pp. 51–58. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006243900510058

[17] S. Adikari and K. Dutta, "Identifying fake profiles in linkedin," in *PACIS*, 2014.

[18] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, 2012, pp. 197–210.

[19] S. Y. Wani, M. M. Kirmani, and S. I. Ansarulla, "Prediction of fake profiles on facebook using supervised machine learning techniques-a theoretical model," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 7, no. 4, pp. 1735–1738, 2016.

[20] M. Albayati and A. Altamimi, "MDFP: A Machine Learning Model for Detecting Fake Facebook Profiles Using Supervised and Unsupervised Mining Techniques," *International journal of simulation: systems, science & technology*, Mar. 2020. [Online]. Available: https://edas.info/doi/10.5013/IJSSST.a.20.01.11

[21] S. Khaled, N. El-Tazi, and H. M. O. Mokhtar, "Detecting Fake Accounts on Social Media," in *2018*

IEEE International Conference on Big Data (Big Data). Seattle, WA, USA: IEEE, Dec. 2018, pp. 3672–3681. [Online]. Available: https://ieeexplore.ieee.org/document/8621913/

[22] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of facebook statuses using naive bayes classifier for language learning," in *IISA 2013*. IEEE, 2013, pp. 1–6.

[23] P. D. Zegzhda, E. V. Malyshev, and E. Y. Pavlenko, "The use of an artificial neural network to detect automatically managed accounts in social networks," *Automatic Control and Computer Sciences*, vol. 51, no. 8, pp. 874–880, Dec. 2017. [Online]. Available: http://link.springer.com/10.3103/S0146411617080296

[24] K. Skorniakov, D. Turdakov, and A. Zhabotinsky, "Make social networks clean again: Graph embedding and stacking classifiers for bot detection," in *CIKM Workshops*, 2018.

[25] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *CoRR*, vol. abs/1703.03107, 2017. [Online]. Available: http://arxiv.org/abs/1703.03107

[26] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A new approach to bot detection: Striking the balance between precision and recall," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 533–540.

[27] "MongoDB," 2020. [Online]. Available: https://www.mongodb.com/

[28] "Docker," 2020. [Online]. Available: https://www.docker.com/

[29] "DigitalOcean," 2020. [Online]. Available: https://www.digitalocean.com/

[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in neural information processing systems*, 2018, pp. 6638–6648.

[31] "Join The Web's Largest Proxy Network - Microleaves Proxies," 2020. [Online]. Available: https://microleaves.com/