Lecture Notes in Networks and Systems 381

Tatiana Antipova *Editor*

Digital Science

DSIC 2021

Springer

Creating and Using Synthetic Data for Neural Network Training, Using the Creation of a Neural Network Classifier of Online Social Network User Roles as an Example

A. N. Rabchevskiy[1,2(✉)] and L. N. Yasnitskiy[2,3]

[1] JSC "SEUSLAB", Perm, Russia
ran@psu.ru
[2] Perm State University, Perm, Russia
[3] National Research University "Higher School of Economics", Perm, Russia

Abstract. The use of synthetic data to train neural networks is becoming increasingly popular. Neural network classification of social network user profiles involves collecting large amounts of personal data, which is associated with high costs and the risk of leaks of confidential information. In this article we suggest using synthetic data to train the neural network classifier of social roles of users in online social networks that actively publish various kinds of materials (posts, reposts, comments) in social networks during the most active phase of the political protests, which can reduce the cost of data acquisition and maintain confidentiality of personal user data. Here is an example of dataset creation based on the algorithm that takes into account the ranges of neural network input parameters' values obtained from the analysis of real data and expert knowledge about correlations between values of different parameters for different user roles. Training and testing of the neural network has been carried out in several packages. Neural network classifier validation was done by comparing classification results of the synthetic neural network model with real data from several user samples. The validation results showed the adequacy of the synthetic neural network model to the real data. The effectiveness of dataset synthesis in cases where it is difficult or impossible to obtain the real data has been shown.

Keywords: Synthetic data · Dataset synthesis · Dataset creation · Social network · User roles · Neural networks · Classification

1 Introduction

The development of modern artificial intelligence technology is impossible to imagine without the use of synthetic data. As the name implies, this is data that is created artificially rather than from real events. They are often created by algorithms and are used for a wide range of activities. According to [1, 2], synthetic data is cheap to produce and can be useful for developing artificial intelligence models, deep learning and software testing. Data privacy provided by synthetic data is one of the most important advantages of synthetic data. User data often includes personal information

T. Antipova (Ed.): DSIC 2021, LNNS 381, pp. 412–421, 2022.
https://doi.org/10.1007/978-3-030-93677-8_36

and personal health information, and synthetic data allows companies to build software without disclosing user data to developers or software tools.

Most machine learning models require a lot of data to be more accurate. Synthetic data can be used to increase the size of training data for machine learning models.

Synthetic data generation creates labelled data instances that are ready to be used in training. This reduces the need for labour-intensive labelling efforts.

Over the last few years, generative models based on deep learning have gained increasing interest and offer some surprising improvements in this area [3]. Relying on huge amounts of data, well-designed network architectures and intelligent learning techniques, deep generative models have demonstrated an incredible ability to produce very realistic pieces of content of various types, such as images, texts and sounds.

Although synthetic data first began to be used in the 1990s, the abundance of computing power and storage space in the 2010s led to the wider use of synthetic data.

There is now a whole industry for the production of synthetic data. Basic use cases and software tools for synthetic data production are presented in [4]. It seems that some of these scenarios can be applied to the actual task of creating a neural network classifier for online social network users.

Online social networks are now a significant factor in people's daily lives and can serve both to communicate between users and to influence users to achieve marketing goals or to propagandize, agitate and mobilize people for certain protests and other illegal actions. To counteract such negative influences, social media has been extensively studied in academic circles [5–10], including in-depth analysis to identify structural and informational evidence of purposeful influence on the network. One element of this analysis is to identify the roles that users play in various social phenomena in online social networks. Various methods have been used to identify users' roles. Recently, neural networks have been increasingly used to classify users into roles. Qualitative classification requires a dataset corresponding to the subject area being modelled and an optimal neural network model. However, professionals often face the problem of lack of quality datasets for training, validation and testing of neural networks.

This paper presents an example of creating and using an artificially synthesised dataset to train, test and validate a neural network classifier for the roles of online social network users who actively publish various types of content (posts, reposts, comments) on social networks during the most active phase of political protest actions.

2 Materials

Social roles manifest themselves in different forms of user activity online. Various data can be used to categorise users into classes, including: number and type of publications, user behaviour patterns, etc. The aggregate of such data is interpreted as a conditional user profile. Neural network classification techniques can be used to classify such profiles into groups with similar parameter values.

In particular, [11] proposes a hybrid neural network to classify text in order to detect users' intentions. In [12], the use of deep neural networks to classify the

sentiments of Twitter users is presented. Judging by the dates of these publications, classification of sets using neural networks is becoming increasingly popular.

In order to perform neural network classification of user roles, it is necessary to have high-quality training and validation sets. To solve this problem, the authors of [13] developed a special dataset of 1000 user profiles to be used to train a neural network identifying the social roles of Twitter users. Creating a dataset based on 740 thousand messages is presented in [14], while work [15] used a dataset consisting of more than 1.2 million text messages extracted from an online higher education community in Australia. To train their neural network, the authors [16] created a dataset based on content analysis of 350 million messages on Twitter. The authors of [17, 18] used ready-made datasets to analyze online social networks. The use of off-the-shelf datasets is convenient, but can be associated with both difficulty in obtaining them and incomplete correspondence of the off-the-shelf dataset to the subject area for which it is to be applied. In addition, neural networks trained on some networks may be unsuitable for other networks.

Because the distribution of roles among users in social networks is extremely heterogeneous, a high-quality dataset requires collecting data on several hundred thousand users, which is very expensive and implies a possible risk of leakage of sensitive data. Using an artificially synthesised dataset solves the problem of preserving user privacy and significantly reduces costs.

3 Method

The synthetic dataset was created based on a random data generation algorithm that takes into account the ranges of input parameter values of the neural network derived from the analysis of real social network users' data and expert knowledge about the correlation between the values of different neural network input parameters for different user roles.

3.1 Description of User Roles

Before creating a classifier of user roles in social media, the term 'social role' needs to be defined. The paper [19] suggests standardising the use of the term 'social role' in online communities as a set of social, psychological, structural and behavioural attributes, and proposes strategies for defining the social roles of users in some online communities. However, the authors do not propose a strict classification of users' social roles, as the set and definition of social roles depends on both the type of social community and the context in which user roles are considered.

In creating the expert neural network, we used the following classes of social media users:

1. A poster is an idea generator, a content creator, often an opinion leader, and with a lot of connections can unite many users around him or her.
2. Reposter - a distributor of ideas, rarely creates content, mostly reposts ready-made publications, aims to spread other people's publications as much as possible.

3. Commentator - does not create content, does not repost, but leaves lots of comments, participates in discussions and debates. Often he/she creates superfluous comments to increase the popularity of the topic of discussion.
4. Universal - a member who actively publishes posts, reposts and comments without a clear predominance of any one type of material.
5. Passive participant - a user who is not very active in the network in terms of creating content, reposts or comments, but regularly visits various pages of the social network. He or she is a recipient of all the information created by the Posters, Reposters and Commenters.

3.2 Neural Network Input and Output Parameters

The following parameters were used as input data to classify users:

X1 - Age of the account
X2 - Number of friends
X3 - Number of posts published
X4 - Number of published reposts
X5 - Number of published comments.

The outputs of the neural network model were:

D1 - takes the value 1 if the user is a Poster and 0 if not.
D2 - takes on a value of 1 if the user is a Reporter and 0 if not.
D3 - takes a value of 1 if user is a Commentator and 0 if not.
D4 - takes value 1 if user is a Universal and 0 if not.
D5 - takes value 1 if the user is a Passive Participant and 0 if not.

3.3 Dataset Synthesis

The main task in generating the dataset was to determine the ranges of each of the input parameters for each role and to introduce certain patterns into the dataset. The values needed to generate the dataset were the ranges obtained from the analysis of material published by users of the Vkontakte social network regarding a fake news blast about the existence of the so-called "Putin's Palace" (see Table 1).

In order for the neural network to learn well and to classify the input sets qualitatively, it was necessary to provide enough examples for each role. In our case, 400 examples were generated for each role.

3.4 Set Synthesis Algorithms for Each Role

For the synthesis of the sets for each role, the value ranges and expert parameter ratios corresponding to the subject area presented in Table 1 were used. The generation of the set was done using Microsoft Excel 2016. To generate the set, a random function was used to select a value from the range of values of the specified package. Let us denote this function as

Table 1. Value ranges for each role derived from the analysis of Vkontakte users' postings regarding the so-called "Putin's Palace"s.

Parameter	Poster	Reposter	Commentator	Universal	Passive
X1 - Age of account (days)	311–4382	0–4881	0–4183	545–4553	86–5170
X2 - Number of friends	0–32368259	0–31586803	0–28990781	0–34031059	0–36014728
X3 - Number of posts	2–94	0–94	0–6	0–17	0–1
X4 - Number of reposts	0–7	2–160	1–3	1–19	0–1
X5 - Number of comments	0–7	0–48	2–48	0–11	0–1

$$R(X_{min}; X_{max}) \tag{1}$$

Passive participants are defined as those who do not have a high activity rating, the ratio between the different types of material does not matter, as long as the values are within the maximum and minimum values. Table 2 shows the formulas used to generate the set for the Passive Participant role. The step of decreasing age of the account is denoted by Δ.

Table 2. Formulas for generating a set for the Passive Participant role

N	X1	X2	X3	X4	X5
1	$X1_{max}$	$X2_{max}$	$X3_{max}$	$X4_{max}$	$X5_{max}$
2	$X1_{max}$-Δ	$R(X2_{min}; X2_{max})$	$R(X3_{min}; X3_{max})$	$R(X4_{min}; X4_{max})$	$R(X5_{min}; X5_{max})$
3	$X1_{max}$-2Δ	$R(X2_{min}; X2_{max})$	$R(X3_{min}; X3_{max})$	$R(X4_{min}; X4_{max})$	$R(X5_{min}; X5_{max})$
..	..	..	..	..	..
N	$X1_{min}$	$X2_{min}$	$X3_{min}$	$X4_{min}$	$X5_{min}$

A poster is an active member whose main activity is creating posts. Let

p_i - the number of posts published by user i,
r_i - the number of reposts published by user i,
k_i - the number of comments posted by user i,

then the total number of submissions by this user m_i can be expressed as

$$m_i = p_i + r_i + k_i, \tag{2}$$

According to experts, a Poster is a user who mainly creates content, often reposts and occasionally comments on other users' posts. Thus, a Poster is a user, whose posts constitute at least 60% of all the materials published by the Poster, whose number of

reposts does not exceed 37% and whose number of comments does not exceed 3%. That is, the Poster must meet the following conditions:

$$p_i \geq \alpha m_i, r_i \leq \beta m_i, k_i \leq \gamma m_i, \tag{3}$$

where $\alpha = 0.6$, $\beta = 0.37$, $\gamma = 0.03$. In this case the values of r_i and k_i can be expressed as

$$r_i \leq \delta p_i \text{ and } k_i \leq \varepsilon p_i, \text{ where } \delta = \frac{\beta}{\alpha} \text{ and } \varepsilon = \frac{\gamma}{\alpha}. \tag{4}$$

Using these relations and the values for a given role from Table 1, we present a set of formulas for generating the role Poster (see Table 3).

Table 3. Formulas for generating a set for the Poster role

n	X1	X2	X3	X4	X5
1	$X1_{max}$	$X2_{max}$	$X3_{max}$	R(0; δ X3)	R(0; ε X3)
2	$X1_{max}$-Δ	R($X2_{min}$; $X2_{max}$)	R($X3_{min}$; $X3_{max}$)	R(0; δ X3)	R(0; ε X3)
3	$X1_{max}$-2Δ	R($X2_{min}$; $X2_{max}$)	R($X3_{min}$; $X3_{max}$)	R(0; δ X3)	R(0; ε X3)
..	..	..	..	..	..
N	$X1_{min}$	$X2_{min}$	$X3_{min}$	$X4_{min}$	$X5_{min}$

According to experts, a Reposter is a user who mainly reposts content created by other users, often creates content themselves and occasionally comments on other users' posts. Thus, a Reposter is a user whose number of reposts is at least 60% of all the materials they have published, whose number of posts is no more than 37% and whose number of comments is no more than 3%. In other words, a user has to meet the following conditions:

$$r_i \geq \alpha m_i, p_i \leq \beta m_i, k_i \leq \gamma m_i, \tag{5}$$

where $\alpha = 0.6$, $\beta = 0.37$, $\gamma = 0.03$. In this case the values of p_i and k_i can be expressed as

$$p_i \leq \delta r_i \text{ and } k_i \leq \varepsilon r_i, \text{ where } \delta = \frac{\beta}{\alpha} \text{ and } \varepsilon = \frac{\gamma}{\alpha} \tag{6}$$

The set of formulas for generating the Reposter role is shown in Table 4.

A commenter is a user who mainly comments on other users' posts, does not often repost and rarely creates content himself. Thus, a Commentator is a user who has at least 60% of their comments, no more than 10% of their posts and no more than 30% of their reposts. That is, the user must meet the following conditions:

Table 4. Formulas for generating a set for the Reposter role

n	X1	X2	X3	X4	X5
1	$X1_{max}$	$X2_{max}$	R(0; δ X4)	$X4_{max}$	R(0; ε X4)
2	$X1_{max}$-Δ	R($X2_{min}$; $X2_{max}$)	R(0; δ X4)	R($X4_{min}$; $X4_{max}$)	R(0; ε X4)
3	$X1_{max}$-2Δ	R($X2_{min}$; $X2_{max}$)	R(0; δ X4)	R($X4_{min}$; $X4_{max}$)	R(0; ε X4)
..	..	..	..	..	..
N	$X1_{min}$	$X2_{min}$	$X3_{min}$	$X4_{min}$	$X5_{min}$

$$k_i \geq \alpha m_i, p_i \leq \beta m_i, r_i \leq \gamma m_i, \tag{7}$$

where $\alpha = 0.6$, $\beta = 0.1$, $\gamma = 0.3$. In this case the values of p_i and r_i can be expressed as

$$p_i \leq \delta k_i \text{ and } r_i \leq \varepsilon k_i, \text{ where } \delta = \frac{\beta}{\alpha} \text{ and } \varepsilon = \frac{\gamma}{\alpha}. \tag{8}$$

The set of formulas for generating the Commentator role is shown in Table 5.

Table 5. Formulas for generating a set for the Commentator role

n	X1	X2	X3	X4	X5
1	$X1_{max}$	$X2_{max}$	R(0; δ X5)	R(0; ε X5)	$X5_{max}$
2	$X1_{max}$-Δ	R($X2_{min}$; $X2_{max}$)	R(0; δ X5)	R(0; ε X5)	R($X5_{min}$; $X5_{max}$)
3	$X1_{max}$-2Δ	R($X2_{min}$; $X2_{max}$)	R(0; δ X5)	R(0; ε X5)	R($X5_{min}$; $X5_{max}$)
..	..	..	..	..	..
N	$X1_{min}$	$X2_{min}$	$X3_{min}$	$X4_{min}$	$X5_{min}$

A Universal is a user who does not meet conditions (3, 5, 7). In fact, a Universal is an active user who is not a Poster, Reporter or Commentator. Using this representation and the graphical values for this role from Table 1, we present a set of formulas for generating the set of Universal role (see Table 6).

Table 6. Formulas for generating a set for the Universal role

n	X1	X2	X3	X4	X5
1	$X1_{max}$	$X2_{max}$	$X3_{max}$	$X4_{max}$	$X5_{max}$
2	$X1_{max}$-Δ	R($X2_{min}$; $X2_{max}$)	R($X3_{min}$; $X3_{max}$)	R($X4_{min}$; $X4_{max}$)	R($X5_{min}$; $X5_{max}$)
3	$X1_{max}$-2Δ	R($X2_{min}$; $X2_{max}$)	R($X3_{min}$; $X3_{max}$)	R($X4_{min}$; $X4_{max}$)	R($X5_{min}$; $X5_{max}$)
..	..	..	..	..	..
N	$X1_{min}$	$X2_{min}$	$X3_{min}$	$X4_{min}$	$X5_{min}$

The example sets for each role derived from the results of the algorithm have been combined, mixed and split into two parts:

- Teaching - 1700 examples,
- Test - 300 examples.

The prepared dataset was used for training and testing the neural network model on Neurosimulator 5.0 platform Nsim5sc [20] (access www.LbAi.ru). As a result of numerous iterations, the best result was obtained by a perceptron neural network with five input neurons, one hidden layer with seven neurons and five output neurons. The hyperbolic tangent was used as the activation functions of all neurons. The formula was used to estimate the error in the neural simulator:

$$E = \frac{\sqrt{\frac{\sum_{n=1}^{N}(d_n - y_n)^2}{N}}}{|\max(d_n) - \min(d_n)|} \ 100\%, \tag{9}$$

where N is the number of sample elements, d_n is the declared role of the *n-th* user, and y_n is its role evaluated by the neural network. The testing error of the neural network model of user role classification is presented in Table 7.

Table 7. Test result of a neural network model of a user role classifier based on the Nsim5sc Neural Simulator

No	Role name	Error %
Y1	Poster	10,3%
Y2	Reposter	10,2%
Y3	Commentator	2,9%
Y4	Universal	16,5%
Y5	Passive member	6,3%

In addition, a neural network model based on a synthetic dataset was trained and tested in other neural network packages (TensorFlow, Apple Create ML, Orange Data Mining) and in all cases the neural network with the same hyperparameters showed the best result.

The validation of the neural network model, based on the artificially synthesised dataset, was carried out by classifying real users and then analyzing the identified user roles. The validation was carried out on a sample of users actively posting different types of material (posts, reposts, comments) on social media during the most active phase (10–15 days) of protest political actions related to the fake news blast about the existence of the so-called “Putin’s Palace”. In addition, a validation was carried out on a sample related to protest actions around the 2020 presidential elections in Belarus. The results of the validation showed full coincidence of the classification results with the proposed algorithm and the results of analytical research of the data of real social network users, performed by expert-analysts.

4 Conclusion

The synthetic dataset generation algorithm used parameter value ranges derived from real data analysis and expert knowledge of the relationships between different parameter values for different roles. Using synthetic data to create a neural network classifier of user roles significantly reduced the cost of creating a dataset from real data and eliminated the risk of leakage of confidential data of social network users. At the same time, the neural network model showed low error rate and adequacy to real data of the target area.

Synthetic data for training and testing the neural network model has been registered as a computer database [21] and is available for use by completing the web request form on the website [22].

The use of synthetic datasets is a universal method and can be recommended for use when it is impossible or difficult to obtain real data for a dataset or when data confidentiality is required. In addition, it can be suggested to use this method to increase the number of examples in datasets when the number of real examples is not enough for quality training of a neural network.

References

1. Dilmegani, G.: The Ultimate Guide to Synthetic Data in 2021. https://research.aimultiple.com/synthetic-data/
2. Dilmegani, G.: Synthetic Data Generation: Techniques, Best Practices & Tools. https://research.aimultiple.com/synthetic-data-generation/
3. Lauterbach, A., Bonime-Blanc, A., Bremmer, I.: The Artificial Intelligence Imperative: A Practical Roadmap for Business. ABC-CLIO (2018)
4. Dilmegani, G.: Top 20 Synthetic Data Use Cases & Applications in 2021. https://research.aimultiple.com/synthetic-data-use-cases/
5. Castells, M.: Networks of Outrage and Hope. Social Movements in the Internet Age. Polity, Cambridge (2012)
6. Faris, D.M.: Dissent and Revolution in a Digital Age. I.B.Tauris (2013). https://doi.org/10.5040/9780755607839
7. Gerbaudo, P.: Tweets and the Streets. Social Media and Contemporary Activism. Pluto Books, London (2012)
8. Tindall, D.B.: From metaphors to mechanisms: critical issues in networks and social movements research. Soc. Netw. **29**, 160–168 (2007). https://doi.org/10.1016/j.socnet.2006.07.001
9. Bennett, W.L., Segerberg, A.: The logic of connective action. Inf. Commun. Soc. **15**, 739–768 (2012). https://doi.org/10.1080/1369118X.2012.670661
10. Juris, J.S.: Reflections on #Occupy Everywhere: social media, public space, and emerging logics of aggregation. Am. Ethnol. **39**, 259–279 (2012). https://doi.org/10.1111/j.1548-1425.2012.01362.x
11. Liu, Y., Liu, H., Wong, L.-P., Lee, L.-K., Zhang, H., Hao, T.: A hybrid neural network RBERT-C based on pre-trained RoBERTa and CNN for user intent classification. In: Zhang, H., Zhang, Z., Wu, Z., Hao, T. (eds.) NCAA 2020. CCIS, vol. 1265, pp. 306–319. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-7670-6_26

12. Abdelhade, N., Soliman, T.H.A., Ibrahim, H.M.: Detecting twitter users' opinions of arabic comments during various time episodes via deep neural network. In: Hassanien, A.E., Shaalan, K., Gaber, T., Tolba, M.F. (eds.) AISI 2017. AISC, vol. 639, pp. 232–246. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-64861-3_22
13. Sunghwan, M.K., Stephen, W., Cecile, P.: Detecting social roles in twitter. In: Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media, Austin, TX, pp. 34–40 (2016)
14. Matsumoto, K., Yoshida, M., Kita, K.: Classification of emoji categories from tweet based on deep neural networks. In: Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval - NLPIR 2018, New York, NY, USA, pp. 17–25. ACM Press (2018). https://doi.org/10.1145/3278293.3278306
15. Wijenayake, P., de Silva, D., Alahakoon, D., Kirigeeganage, S.: Automated detection of social roles in online communities using deep learning. In: Proceedings of the 3rd International Conference on Software Engineering and Information Management, New York, NY, USA, pp. 63–68. ACM (2020). https://doi.org/10.1145/3378936.3378973
16. Lin, H., et al.: User-level psychological stress detection from social media using deep neural network. In: Proceedings of the 22nd ACM International Conference on Multimedia, New York, NY, USA, pp. 507–516. ACM (2014). https://doi.org/10.1145/2647868.2654945
17. Jabłońska, M.R., Zajdel, R.: Artificial neural networks for predicting social comparison effects among female Instagram users. PLoS ONE **15** (2020). https://doi.org/10.1371/journal.pone.0229354
18. Segalin, C., et al.: What your facebook profile picture reveals about your personality. In: Proceedings of the 25th ACM International Conference on Multimedia, New York, NY, USA, pp. 460–468. ACM (2017). https://doi.org/10.1145/3123266.3123331
19. Gleave, E., Welser, H.T., Lento, T.M., Smith, M.A.: A conceptual and operational definition of "social role" in online community. In: 2009 42nd Hawaii International Conference on System Sciences. IEEE (2009). https://doi.org/10.1109/HICSS.2009.6
20. Cherepanov, F.M., Yasnitsky, L.N.: Neurosimulator 5.0: Rospatent Certificate of State Registration of Computer Programme No. 2014618208 dated 12.07.2014
21. Rabchevskiy, A.N., Zayakin, V.S.: A database for the classification of roles of social network users. State Registration Certificate for the Computer Database No. 2021621533 dated 15.07.2021
22. https://seuslab.ru/registered_db/2021621533?lang=en