# A Multimodal Approach to Synthetic Personal Data Generation with Mixed Modelling: Bayesian Networks, GAN's and Classification Models

Irina Deeva(✉), Andrey Mossyayev, and Anna V. Kalyuzhnaya

ITMO University, Saint-Petersburg, Russia

**Abstract.** Personal data is multimodal, as it is represented by various types of data - tabular data, images, text data. In this regard, the generation of synthetic personal data requires a large number of interconnected datasets, but it is often very difficult to collect tabular data, images or texts for the same people. The problem of having interconnected datasets can be solved by separating the models to generate each type of data and combining them into a single model pipeline. This paper presents a multimodal approach to generating synthetic personal data of a social network user, which allows generating socio-demographic information in the user's profile (tabular data), an image of the user's avatar and content images that correlates with the user's interests. The multimodal approach is based on the combined use of Bayesian networks, generative adversarial networks and discriminative model. This approach, due to the independent training of models, allows us to solve the problem of the presence of interconnected data sets (info + photos) and can also be used for example to anonymize medical data. A quantitative assessment shows that the obtained synthetic profiles are quite plausible.

**Keywords:** Synthetic personal data · Bayesian networks · Generative adversarial networks · Multimodal approach · Classification models

## 1 Introduction

Dataset availability is a critical factor in the development of artificial intelligence and machine learning projects. More broadly, data is needed to train and test machine learning models and evaluate already developed applications. The McKinsey Global Institute points out that data access is one of the main challenges hindering the ubiquity of machine learning projects [6]. Deloitte's analysis showed that data access problems are among the top three problems that companies face when implementing artificial intelligence projects [13]. If we are talking about personal data, then the main reason for the inaccessibility of such data is the presence of confidentiality restrictions.


T. Hara and H. Yamaguchi (Eds.): MobiQuitous 2021, LNICST 419, pp. 847–859, 2022.
https://doi.org/10.1007/978-3-030-94822-1_55

For example, medical data also contains personal information. In this case, we are talking about the need to anonymise data. Many companies facing these challenges prefer to use open-source datasets. However, the main drawback of such datasets is the lack of diversity and the specificity of the composition, which is associated with the task for which the dataset was collected. Data synthesis can solve the problems listed above and enable analysts to work with diverse and realistic data. Synthetic data is not personal data, which means that privacy restrictions do not apply to them.

Since personal data often contain different types of data, such as tabular data and images, it is multimodal (Fig. 1). And in this case, to train the generation model, you need interconnected data, which is rather difficult to collect. This article presents an approach that combines multiple models to generate each data type. The main advantage of this approach is that each model is trained separately on its own set of data, and their combination allows, as a result, to obtain synthetic multimodal data. In the article, all experiments were carried out on the example of generating a user profile of a social network, however, this approach can be extended to any area.

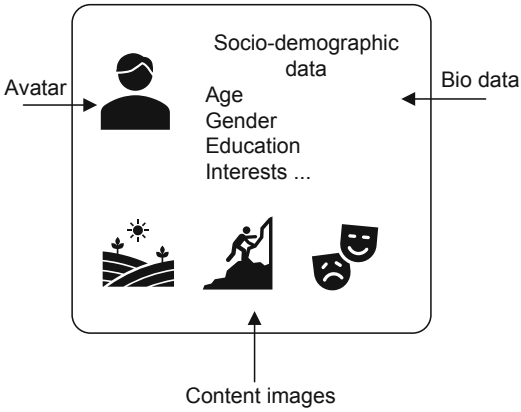



**Fig. 1.** Personal profile as a multimodal object for generation.

## 2 Related Work

This section presents existing approaches for modelling and generating synthetic quantitative and categorical variables and synthetic images.

Personal data generation and further analysis of its properties have long been used in many industries, such as modelling the dynamics of transmission of infectious diseases in American Samoa [28], modelling demographics of households [10], building and testing information discovery systems [17], modelling the behaviour of social media users [23], etc. However, to create synthetic profiles

of people, it is required to generate not only socio-demographic data but also graphic data (an image of a person, pictures of interest).

Various algorithms and methods have been proposed over the past 20 years to generate synthetic population. Standard techniques for synthetic data generation methods can be divided into two categories, corresponding to two different approaches: the first, called Synthetic Reconstruction (SR), aims to create a vector of traits for each entity [27]. The second method, called Combinatorial Optimization (CO), consists of duplicating known real individual data records [26]. Also, there are other proposed methods as Copula-Based Population Generation, which used to understand the dependency structures among different distributions [14]. Bayesian networks are also one of the methods for generating synthetic data and are used to generate synthetic personal data [8]. However, all of the above approaches generate only distributions of numerical data.

Synthetic images generation is one of the most rapidly developing areas. Depending on the architecture, auxiliary image data, there are different kinds of generators. Additional information for generation may serve noise, conditional data (such as classes), text or another image. Bright representative of image generator from noise is Generative Adversarial Network (GAN) [11]. It consists of two competitive networks, Generator and Discriminator, where Generator objective is to output realistic images trying to fool Discriminator. Incorporating side information into the process of generation, researchers came up with Conditional GAN (CGAN) [18]. Depending on the scheme of how additional information fed into models, the architectures like ACGAN [20], cGANs with Projection Discriminator [19] were proposed. The class of models for Image-to-Image generation, which could be used for a generation and Variational Autoencoders (VAE) [16]. VAEs work differently because their objective is to generate such images which distributed as close as possible to the distribution of real images. Indeed, pure generative models still suffering from different kinds of artefacts and occlusions and sometimes cannot deal with complex conditions. For that case, discriminative models have to be used. The new generation of ones utilizes an unsupervised approach for the task, called zero-shot classification. An example of such a model is CLIP [21], where there is an opportunity to map images and texts into the same dimensions, resulting from the probability distribution of the image's caption being described by the text.

The difficulty is the generation of cross-modal profiles, in which data with different nature are aggregated (for example, posts and photos on social networks, demographic data). In [15], approaches based on autoencoders for the simultaneous generation of text and images, or the restoration of missing parts, are proposed. However, such an approach requires many associated tagged data, which is not always available, especially if we try to generate faces and socio-demographic information. Our approach, based on combining Bayesian networks for generating socio-demographic information, GANs for generating a portrait and a zero-shot classification model for getting the content images, allows us to train each model on each dataset separately. Also, such a separation will enable you to get rid of a significant bias in the data when, for example, a portrait of

a black person is not generated for the characteristics of a person with higher education. The segregated training that we offer is free from such mistakes. And as you know, the tolerance of artificial intelligence is a big problem [29,30].

## 3 Method

The multimodal method for generating synthetic personal data consists of the sequential use of three models: Bayesian network, InterFaceGAN and classifier model.

### 3.1 Tabular Data Generation Model

A Bayesian network is used to generate socio-demographic data. Bayesian network is an oriented probabilistic model that allows you to reduce the size of the original multidimensional data due to the rule of conditional independence. Thus, having a set of real data, you can train a Bayesian network on them and generate synthetic data by sampling from the trained network. The simplest algorithm for learning the structure of a network is the greedy Hill-Climbing algorithm, the essence of which is to find a structure that maximizes the scoring function [5,9]. The likelihood or, for example, the K2 function [7] can be chosen as a score function. For learning the parameters of distributions at network nodes in the presence of both continuous data and discrete data, it is common practice to discretize continuous data, but this leads to information loss. In our previous studies [4], we investigated a way to learn distribution parameters in nodes on mixed data without discretization by using conditional Gaussian distributions. Therefore, it does not result in a loss of quality and will be used in this study. Forward sampling from the Bayesian network is used to generate synthetic data [12].

### 3.2 Faces Images Generation Model

To generate a synthetic personality portrait, it was necessary to take into account the fact that the appearance of a synthetic personality must be combined with the generated biographical data. Therefore generation strategy was taken from InterFaceGAN [24], i.e. images were generated in an unconditional manner from noise, then passed to auxiliary classifiers (Fig. 2). In our case, the classification of portraits by gender, age and ethnicity was chosen, since it is these characteristics that most clearly define a person's appearance. These classifiers were trained with Logistic Regression on top of the embeddings. Embeddings retrieved using Dlib [2] package on UTKFace dataset [3]. The dataset consists of 20k face images in the wild with all needed parameters labelled.

### 3.3 Content Images Generation Model

Except for the portrait images, it is natural for users to post ones of nature, pets, cars etc. Indeed, the distribution of topics of such images is complex and described by various social environments and circumstances. For simplicity, we decided to condition content images entirely on topics of user's interest. That is shown in Fig. 3, where keywords taken from the top 3 most probable topics of user interest, then mentioned keywords have to be pre-processed, following recommendations from [21] and added one placeholder class whose aim is to spread the probabilities and make filtering more granular. Coupling with source images from [1], they passed to CLIP [21] model, resulting in "conjugate" matrix, where elements correspond to probability image having concrete text description. This approach rather filtering and discriminative than generative. Nevertheless, it works in an unsupervised manner and grants flexibility and trustworthiness; in this case, some of the topics' keywords might change. We recalculate text embeddings and get a new resulting matrix.

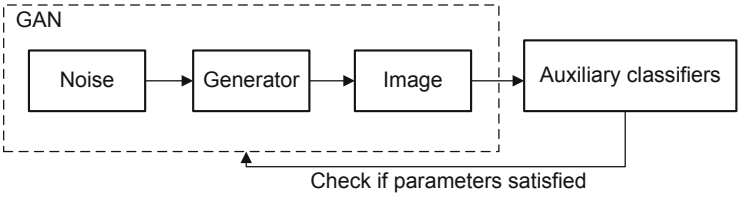



**Fig. 2.** Conditional image generation pipeline.

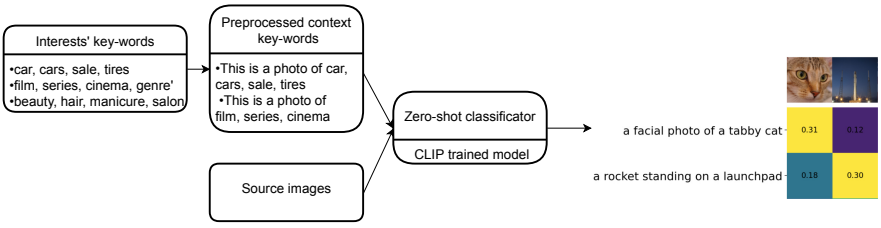



**Fig. 3.** Profile content images generation pipeline.

### 3.4 Multimodal Data Generation approach

The proposed multimodal method consists of sequentially running the following steps:

1. A user sets the criteria for a synthetic personality (age and gender). If the criteria are not set, then random personality will be generated;
2. The function of sampling from the Bayesian network is started with the parameters set at the first step;
3. The name (ethnicity) is randomly assigned to the generated personal data;
4. The generated synthetic personal data is fed to the GAN input, which generates a portrait for each synthetic personality;
5. A textual description of the first most likely interests of a person is fed to the input of the classifier model, and three pictures are selected for the synthetic profile;
6. The result is displayed as an example of synthetic profiles.

The main advantage of this approach is that each pipeline model can be trained separately on its own data, which does not require the presence of an associated dataset (information + images). Also, the peculiarity of the method is that information about the user is generated separately, and in this generation, there is no link to nationality since the node with the ethnicity is not connected with the rest of the nodes of the Bayesian network. This was done so that the resulting profiles were more diverse; for example, there was a sufficient number of black people with higher education. In this way, we try to reproduce rare combinations of characteristics in sufficient numbers to increase diversity. The model for generating faces is also trained on its face database. Such independent learning allows us to generate a portrait of a synthetic person that is not subject to bias that occurs in interconnected datasets (for example, where only white people have profiles of people with higher education). Figure 4 illustrates the general scheme of the developed service for generating synthetic personal data, which implements the described method.

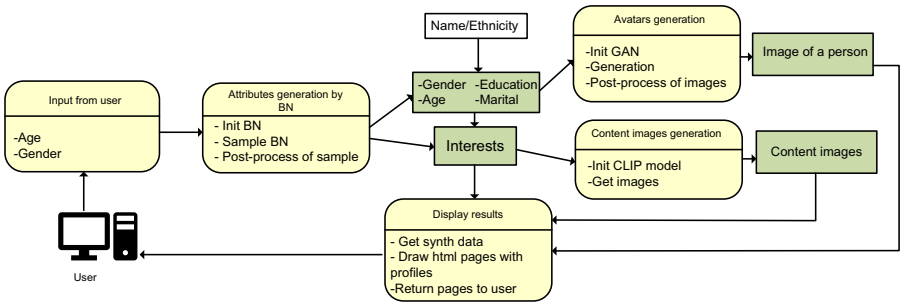



**Fig. 4.** Pipeline of the service for generating multimodal synthetic personal data.

## 4 Experiments Results

### 4.1 Generating Common Datasets

A Bayesian network was first learned on data from a social network. The dataset has 30000 users. Such fields as name, gender, age, higher education, marital status and vector of interests were allocated for experiments. User interests were obtained using the Additive regularization model for topic modelling process (ARTM), and its program implementation BigARTM for Python programming language, running on user subscriptions (groups) [22,25]. In total, 26 different interests were identified, which are described by keywords (5–10 words for each interest). It should be noted that now the interests are highlighted and generated in their raw form, that is, in the form of keywords that describe them. This is necessary for the further development of the profile since the generation of media content of a synthetic user will be added, which will rely on the vector of interests that is now obtained.

Figure 5 shows the structure of the resulting Bayesian network, which was used for tabular data generation. Figure 6, Fig. 7 and Fig. 8 illustrate the results of generating synthetic data in comparison with real data.

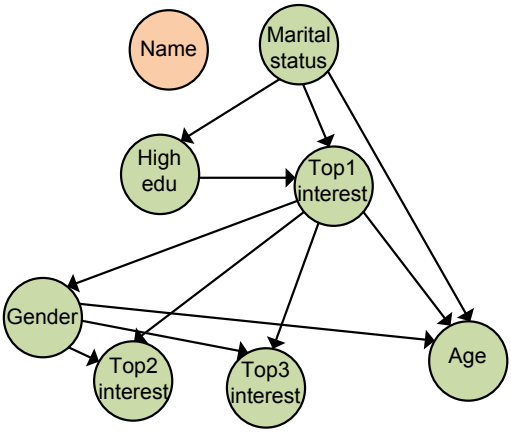
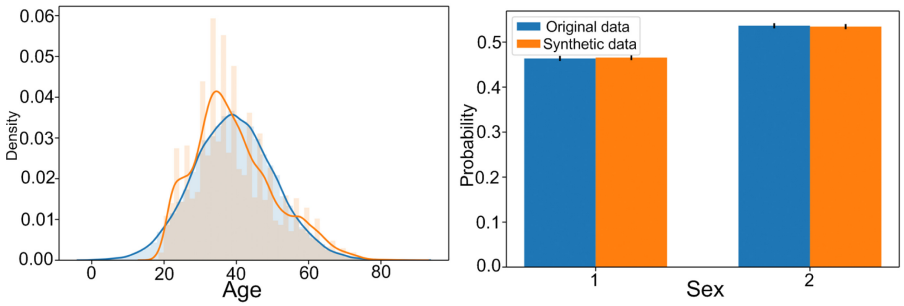



**Fig. 5.** Bayesian network structure.

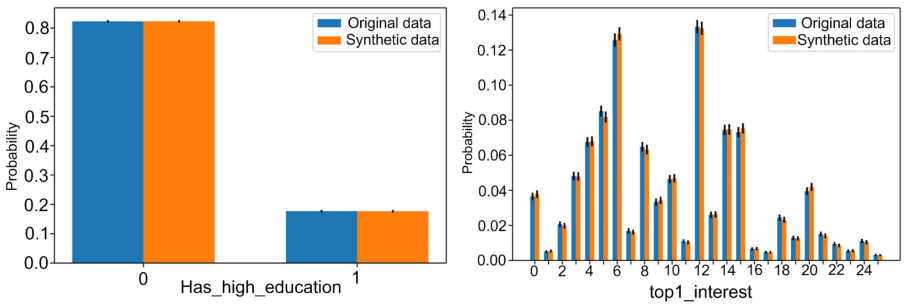
The quality of the obtained synthetic data was assessed through the accuracy of the classifier model. A logistic regression model was trained on data where half was real data (labelled "1"), and a half was synthetic data (labelled "0"). Then the accuracy was checked on a test sample and showed the value of the ROC-AUC metric equal to 0.49. This suggests that the classifier does not distinguish between real and synthetic data, which means they are quite similar.

Then, three fields from the synthetic dataset (age, gender, name) are passed to the face generator's input. The accuracy of face generation by input characteristics has been measured. Table 1 shows the quality of persons' generation



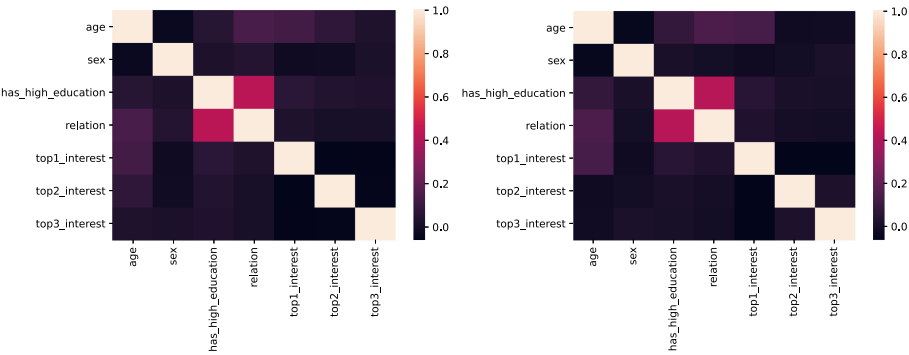

(a) (b)

**Fig. 6.** Comparison of marginal distributions of the original data and synthetic data for age and sex.




(a) (b)

**Fig. 7.** Comparison of marginal distributions of the original data and synthetic data for education and top1 interests.




(a) (b)

**Fig. 8.** Correlation matrices based on real data (a) and synthetic data (b).

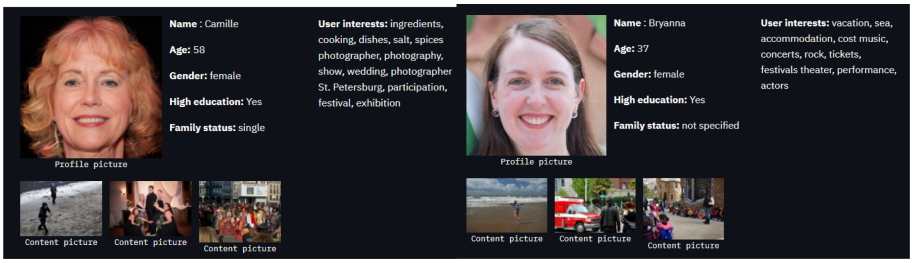
according to the given characteristics of ethnicity (determined by name), age and gender. Also, a textual description of the first three most likely interests was submitted to the input of the discriminator model, and three pictures suitable for the topic were selected.

**Table 1.** Accuracy score for faces generation task.

| Attribute | Precision | Recall | F1-score |
|---|---|---|---|
| Generation by ethnicity | | | |
| White | 0.86 | 0.92 | 0.89 |
| Black | 0.89 | 0.9 | 0.89 |
| Asian | 0.91 | 0.95 | 0.93 |
| Indian | 0.74 | 0.8 | 0.77 |
| Generation by age | | | |
| Teen (15–20) | 0.9 | 0.83 | 0.87 |
| Adult (21–45) | 0.85 | 0.91 | 0.88 |
| Old (46–90) | 0.84 | 0.77 | 0.8 |
| Generation by gender | | | |
| Male | 0.93 | 0.93 | 0.93 |
| Female | 0.93 | 0.93 | 0.93 |

Examples of the obtained synthetic profiles are shown in Fig. 9, Fig. 10 and Fig. 11. The profile displays a synthetic portrait, socio-demographic information, interests and content images.

To assess the credibility of the resulting profiles, we surveyed 80 random people. People were given ten synthetic profiles and estimated whether they believe that this profile belongs to a real person. As a result, 70% of people believed in the reality of the shown synthetic profiles.



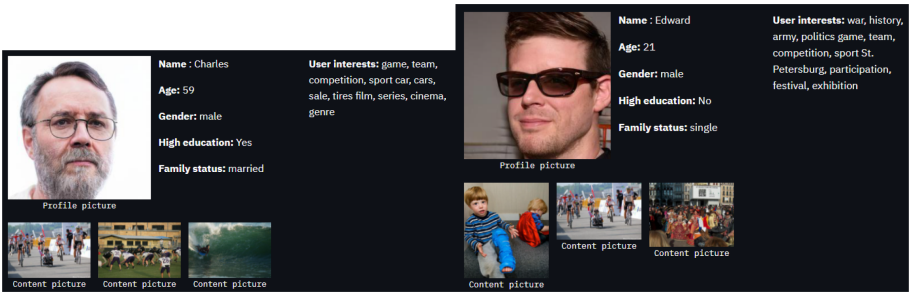

(a) (b)

**Fig. 9.** Examples of synthetic profiles of people with different characteristics (women).

### 4.2 Generation of Specific Populations

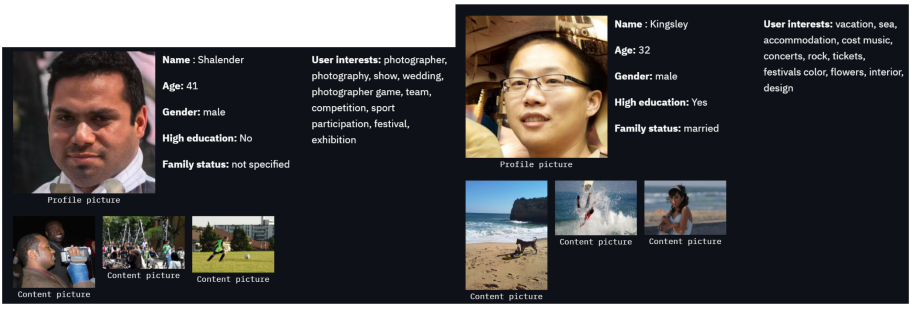
To test the possibility of generating by our service not only common datasets but also populations of people, combined according to a number of characteristics, the following experiments were carried out. At first, users were identified by marital status. From the entire dataset, people with the marital status "married" were selected, then the distribution of the "top 1 interest" for this group of people was built. To generate synthetic data for people with this marital status, the "relation" node was initialized to "married" and sampled. Figure 12a shows the distribution of the "top 1 interest" field for real data and sampling. It can be seen that for this group of people, the most likely interest is interest with code 18 - 'school, education, question, topic', which is also reproduced in the synthetic dataset.

An experiment was also conducted to generate data for a specific gender group. All women were selected from the dataset, and the distribution of the "top 3 interest" parameter was built for them. Then, on the network, the gender node was initialized to "female", and sampling was performed. The figure shows a comparison of the resulting interests. It can be seen in Fig. 12b that in real data



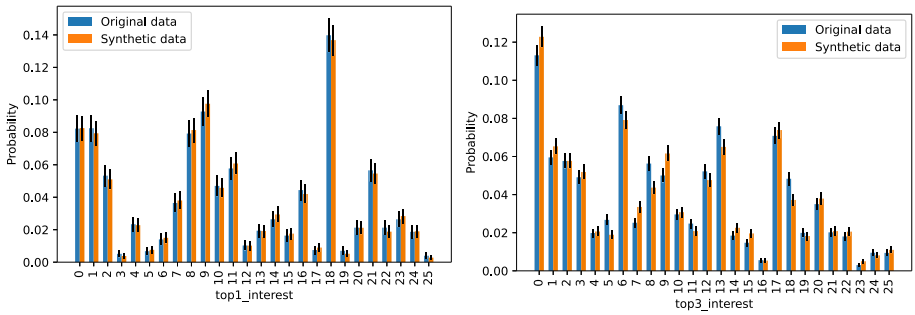

(a) (b)

**Fig. 10.** Examples of synthetic profiles of people with different characteristics (men).




(a) (b)

**Fig. 11.** Examples of synthetic profiles of people with different ethnitics.




(a) (b)

**Fig. 12.** Distribution of real and synthetic data for two social groups. a - distribution for top1 interest for married people. b - distribution for top3 interest for women.

for the gender group "women", the most probable value of the "top 3 interest" field is the interest with the code 0 - 'beauty, hair, manicure, salon', which is reproduced in the synthetic data.

## 5 Conclusion

This paper presented a multimodal approach for generating synthetic personal data. The method based on the joint use of Bayesian networks, GANs and classification model allows independent training of models, which solves the problems of the presence of associated marked datasets (for example, biographical data + photos), and also increases the variety of the resulting profiles, since the incoherence of the datasets on which the models are trained is deprived bias and can generate a variety of profiles. Accuracy measurements have shown that the resulting profiles are quite believable. Experiments also show that such an approach can be used to model data from certain social groups. In the future, it is planned to add additional fields characterizing a person to the synthetic profile and add synthetic text messages. This addition will be based on the vector of interests, which is now being generated in the form of keywords describing the interest. This form will be convenient to use for generating posts. It would also be interesting to consider the possibility of generating rare combinations of characteristics of people as a task of modelling distribution tails to increase the diversity of existing social datasets.

**Acknowledgement.** The reported study was funded by RFBR, project number 20-37-90117.

## References


1. Flickr8k dataset (2020). https://www.kaggle.com/adityajn105/flickr8k
2. Library for embeddings (2020). http://dlib.net/
3. Dataset of faces (2021). https://susanqq.github.io/UTKFace/
4. Bubnova, A.V., Deeva, I., Kalyuzhnaya, A.V.: MIxBN: library for learning Bayesian networks from mixed data. arXiv preprint arXiv:2106.13194 (2021)
5. Chickering, D.M.: Optimal structure identification with greedy search. J. Mach. Learn. Res. **3**(Nov), 507–554 (2002)
6. Chui, M.: Artificial intelligence the next digital frontier, vol. 47, pp. 3–6. McKinsey and Company Global Institute (2017)
7. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
8. Deeva, I., Andriushchenko, P.D., Kalyuzhnaya, A.V., Boukhanovsky, A.V.: Bayesian networks-based personal data synthesis. In: Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good, pp. 6–11 (2020)
9. Gámez, J.A., Mateo, J.L., Puerta, J.M.: Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Min. Knowl. Disc. **22**(1), 106–148 (2011)
10. Geard, N., McCaw, J.M., Dorin, A., Korb, K.B., McVernon, J.: Synthetic population dynamics: a model of household demography. J. Artif. Soc. Soc. Simul. **16**(1), 8 (2013)
11. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv:1406.2661 (2014)
12. Guo, H., Hsu, W.: A survey of algorithms for real-time Bayesian network inference. In: Join Workshop on Real Time Decision Support and Diagnosis Systems (2002)
13. Insights, D.: State of AI in the enterprise (2018)
14. Jeong, B., Lee, W., Kim, D.S., Shin, H.: Copula-based approach to synthetic population generation. PLoS ONE **11**(8), e0159496 (2016)
15. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
16. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. CoRR arXiv:1312.6114 (2014)
17. Lin, P.J., et al.: Development of a synthetic data set generator for building and testing information discovery systems. In: Third International Conference on Information Technology: New Generations (ITNG'06), pp. 707–712. IEEE (2006)
18. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014)
19. Miyato, T., Koyama, M.: cGANs with projection discriminator. arXiv:1802.05637 (2018)
20. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: ICML (2017)
21. Radford, A., et al.: Learning transferable visual models from natural language supervision. arXiv:2103.00020 (2021)
22. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer (2010)
23. Sagduyu, Y.E., Grushin, A., Shi, Y.: Synthetic social media data generation. IEEE Trans. Comput. Soc. Syst. **5**(3), 605–620 (2018)

24. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of GANs for semantic face editing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9240–9249 (2020)
25. Uteuov, A.: Topic model for online communities' interests prediction. Procedia Comput. Sci. **156**, 204–213 (2019)
26. Williamson, P., Birkin, M., Rees, P.H.: The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environ. Plann. A **30**(5), 785–816 (1998)
27. Wilson, A.G., Pownall, C.E.: A new representation of the urban system for modelling and for the study of micro-level interdependence. Area **8**, 246–254 (1976)
28. Xu, Z., Glass, K., Lau, C.L., Geard, N., Graves, P., Clements, A.: A synthetic population for modelling the dynamics of infectious disease transmission in American Samoa. Sci. Rep. **7**(1), 1–9 (2017)
29. Zuiderveen Borgesius, F., et al.: Discrimination, artificial intelligence, and algorithmic decision-making (2018)
30. Zuiderveen Borgesius, F.J.: Strengthening legal protection against discrimination by algorithms and artificial intelligence. Int. J. Hum. Rights **24**(10), 1572–1593 (2020)