

Towards Synthetic Social Media Data

Justina MANDRAVICKAITĖ^{1,2}, Milita SONGAILAITĖ^{1,2}, Veronika
GVOZDOVAITĖ^{1,3}, Danguolė KALINAUSKAITĖ^{1,2}, Tomas KRILAVIČIUS^{1,2}

¹ Vytautas Magnus University, Kaunas, Lithuania

² Centre for Applied Research and Development (CARD), Lithuania

³ University of Oxford, UK

`justina.mandravickaite@vdu.lt`, `milita.songailaite@stud.vdu.lt`,
`gvozdovaite.veronika@gmail.com`, `danguole.kalinauskaite@vdu.lt`,
`tomas.krilavicius@vdu.lt`

Abstract. There is an increasing need for training and testing data that can be used for the development of technologies and research. Due to data protection regulations, a lot of the real-world data – especially the data from social media – becomes unavailable to use. The problem can be solved by generating synthetic data that imitates the properties of real-world data. In this paper, we present *Fabulator* – a social media generator that combines text and graph structure to be used for the generation of synthetic data and, in the future, for the simulation of various events on social media.

Keywords: social media, NLG, synthetic data, SNS, English

1 Introduction

The use of social media data in scientific research is rapidly increasing, typically focusing on events of interest and/or spatially mapping variables of interest. However, concerns for privacy, data management, as well as the realization of General Data Protection Regulation (GDPR) directive have made it more difficult to obtain these data. As real data is unavailable or takes too much effort and other resources to obtain, data synthesis is proposed as a solution for overcoming the shortage of data.

So far, the discussion on social media data generation has been focused on graph structures (Li et al., 2018; You et al., 2018; Grover et al., 2019; Bonifati et al., 2020; Nettleton et al., 2021). Synthesizers of synthetic graphs imitate the structures of real online social networks, therefore making the generated data similar to real social media interactions and reliable for usage in further research. However, these abstract structures do not reach the level of textual representation, and thus do not generate textual data.

Textual data generation on its own has gained a lot of attention in the artificial intelligence community. Perhaps the most notable models are the Generative Pre-trained

Transformers, the latest of which, GPT-3, was released in 2020 by OpenAI. It has produced impressive results in text generation and has even participated in a Reddit social network thread (Ma and Lalor, 2020). Beside GPT, some of the most advanced methods for text generation include BART (Lewis et al., 2019) and other GAN-based (*Generative Adversarial Networks*) approaches (Iqbal and Qureshi, 2020).

Conversation generation, also referred to as response generation, focuses on computer-generated responses to short snippets of real user-generated text. However, even the short-text models that only generate one round of conversation still lack semantic control, i. e., they often generate responses that are semantically irrelevant to the conversation, or even ‘hallucinate’ facts, that is, include false information (Wu et al., 2020).

A successful social media data generator would combine text generation with a structural (graph) generator to get the most reliable data. There is a lack of such solutions. At the moment, the only generator of this kind is SHIELD (Synthetic High Fidelity Social Media Data Generator) (Sagduyu et al., 2018). However, no information has been found on the later use of this generator for research or other purposes. Moreover, the limitation of SHIELD comes from the type of data that was used for its training. It was trained (and evaluated) on the posts from a single platform – Twitter, which makes use of short and relatively simple 140-character tweets.

The aim of our project is to create a prototype for a synthetic social media data generator. Our prototype, the Fabulator, combines the use of graph structures and text generation to produce synthetic data to overcome the shortage of necessary data.

2 System description

2.1 Text generation

For text generation, we took pretrained dialogue response generation model (DialoGPT-medium) and retrained it with 2 different datasets, so the generation of conversations following different topics and perspectives would be possible. We used relevant Reddit data for this retraining which resulted in 2 dialogue response generation models ("political" and "conspiratorial") which interact with each other.

2.1.1 Data

As the goal was the generation of dialogues that were specialized thematically, we used DialoGPT-medium as a base model and Reddit conversations covering politics and conspiracy theories in order to train 2 models capable of generating dialogues and interacting with each other. Our data for training consisted of posts and comments from 5 subreddits and comments from 18 posts, covering the USA political environment and hand-picked by an expert. The decision to focus on it was made because of the amount of data available for retraining models in terms of posts, comments, and conversation strings in general, as well as popularity due to the impact the political events, views, and opinions have on the rest of the world.

Our training data covers discussions and conversations regarding the main political parties (the Republican Party and the Democrat Party) and 4 persons representing

them in the light of the last presidential elections (Joe Biden, Kamala Harris, Donald Trump, and Mike Pence). The latter dataset was used for training the "political" dialogue response generation model. For the training of the second conversational model, we included conversations for and against conspiracy theories, QAnon in particular, as it gained significant popularity in recent months in Reddit mega-threads, in the dataset. QAnon is a conspiracy theory that evolved into a political movement and it came from the American far-right political sphere (Amarasingam and Argentino, 2020; Moskalenko and McCauley, 2021). We also included a small part of random comments into our training datasets to increase the variety of the data. All dialogue strings in our datasets for retraining the selected base model contain at least 4 questions/answers from two or more different users. The detailed statistics of composition of training data presented in Table1 and Table2.

Table 1. Statistics of subreddits used for training.

Subreddit	No. of comments	No. of posts	No. of sentences	No. of words
QAnonCasualties	329965	9958	1070496	15464376
QanonKaren	17549	2753	59472	859132
ParlerWatch	390516	12643	1264982	18302088
Qult_Headquarters	565342	32795	2066057	29382314
Conservative	4168766	273975	1858265	27082264

Table 2. Statistics of selected posts which comments were used for training.

Topic	No. of posts	No. of comments	No. of sentences	No. of words
Joe Biden	10	8248	80994	1221030
Donald Trump	10	5252	17923	295279
Mike Pence	6	117400	67202	1009583
Kamala Harris	6	104900	98290	1488827
the Republican Party	3	9100	19224	31254
the Democrat Party	3	9500	17923	295279
Random	7	5159	59342	829572

2.1.2 Methods

DialoGPT-medium⁴ was selected as a base model. DialoGPT is a state-of-the-art pretrained dialogue generation model for multi-response conversations (Zhang et al., 2019). It is trained originally on Reddit data – multi-response dialogue from the discussion thread. We retrained this base model to generate "political" and "conspiratorial"

⁴ <https://huggingface.co/microsoft/DialoGPT-medium>

conversations. Training of our 2 dialogue models ("political" and "conspiratorial") took 22 hours.

Our models are relatively flexible in terms of the length of generated text. Strings that repeat themselves endlessly during generation, i.e., the same word, phrase, or sentence generated repeatedly, are automatically filtered out if the text becomes too repetitive. Finally, dialogue strings can be generated with different 'personas', i.e. the dialogue can take place between fake speakers following different ideologies as the dialogue snippet in Figure 1 shows.

2.2 Network structure generation

For the graph generation part, we chose to apply numbers of different events and links to be generated randomly by *fakesocial*⁵ social network generator. It generates a simple social network consisting of fake users who have connections, make posts, comment, and like these posts as well. User profiles utilize generated images as profile pictures while generated text serves as posts and comments. The whole social network is packaged as a website.

2.2.1 Fake user profile generation

For the generation of the profile, a data file with new events (posts, comments, and likes) is added based on the existing events. A user profile consists of a name, profile picture, job title, company, and location (city). The profile pictures were taken from the *This Person Does Not Exist*⁶ website. StyleGan (Karras et al., 2020) – data-driven unconditional generative image model – was used to generate them. Markovify⁷ (Markov chain generator) was used with a dataset of job titles⁸ to generate new ones for the fake users. The rest of the profile data was generated by randomly sampling and shuffling data from various lists (company⁹, location (city)¹⁰ and personal names¹¹). An example of a generated user profile, already realized in a fake social network, is presented in Figure 2.

2.2.2 Graph generation

At this stage of the development of Fabulator, the focus was on text generation, so we used a simple fake social network generator *fakesocial* for the generation of network structure. For each generated profile a random number of events, that is, links

⁵ <https://github.com/berfr/fakesocial>

⁶ <https://thispersondoesnotexist.com/>

⁷ <https://github.com/jsvine/markovify>

⁸ <https://github.com/jneidel/job-titles>

⁹ <https://datahub.io/core/nasdaq-listings>

¹⁰ <https://datahub.io/core/world-cities>

¹¹ <https://github.com/treyhunner/names>

```
[
  {
    "Bot": "qanon",
    "Chat": "Biden is going to beat Joe Biden! LoL He is doing such
a great job of this..."
  },
  {
    "Bot": "conservative",
    "Chat": "I mean this cant seriously be a real candidate for POTUS."
  },
  {
    "Bot": "qanon",
    "Chat": "I think youre right that this is not over. There will be
more and more excuses, theories, fear mongering, memes, and hatred."
  },
  {
    "Bot": "conservative",
    "Chat": "I think its also relevant to note that the policies
non-conservatives in power are flirting with are just simple
governance, like voting rights and managing public health crises."
  },
  {
    "Bot": "qanon",
    "Chat": "I think its"
  },
  {
    "Bot": "conservative",
    "Chat": "I'm honestly trying to envision what a 'Biden jacket'
would even be, and drawing a blank."
  },
  {
    "Bot": "qanon",
    "Chat": "She has a nice clothing line to peruse. The only
conservative thing she has going for her is the fact that she
is Jewish, so of course she is supporting Israel. I thought
she was Antifa/Crisis Actor?"
  }
]
```

Fig. 1. Snippet of generated dialogue.

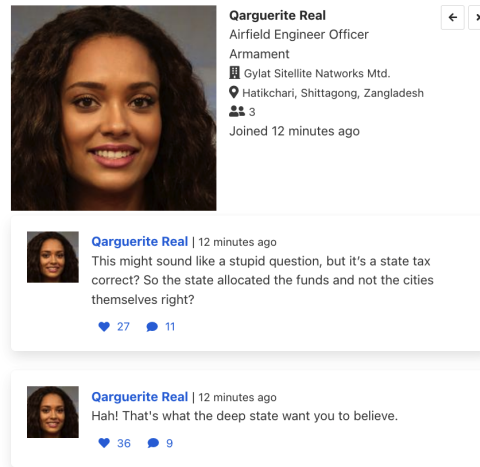


Fig. 2. Snippet of the generated user profile.

and types of interaction (posts, comments, and likes) is assigned, linking all fake users in a network¹².

At the moment generated graph structures are exemplary and play a role of a proof of concept, that is, that generation of social media data, combining text and network structure can be implemented, as we saw the need as well as the lack in terms of such kind of data generator. Therefore the generation of prototype profiles (user profiles covering real-life statistics in terms of attributes, links, and interactions) and graph features corresponding to a particular social networking site or sites is in our future plans.

2.3 Combining synthetic graph and text

The first step in Fabulator is the generation of the desired number of conversation(s) of randomly assigned length(s). Also, for response generation, a mixture of conversational models or a particular model is chosen (e.g., conversations can be generated using the "political" or the "conspiratorial" model or both of them at the same time, allowing interaction between fake speakers with different perspectives). Then fake user profiles are generated. After that, numbers of events of different types (posts, comments, likes) are specified and conversational strings are randomly assigned to each of the fake user profiles based on that specification. Once the specified number of events is realized, a static website with generated fake social network is produced. The snippets of synthetic social network where graph and text components are combined are presented in Figures 3 and 4.

¹² As the basis, we used Facebook dataset, available at Kaggle (<https://www.kaggle.com/datasets/sheenabatra/facebook-data>).

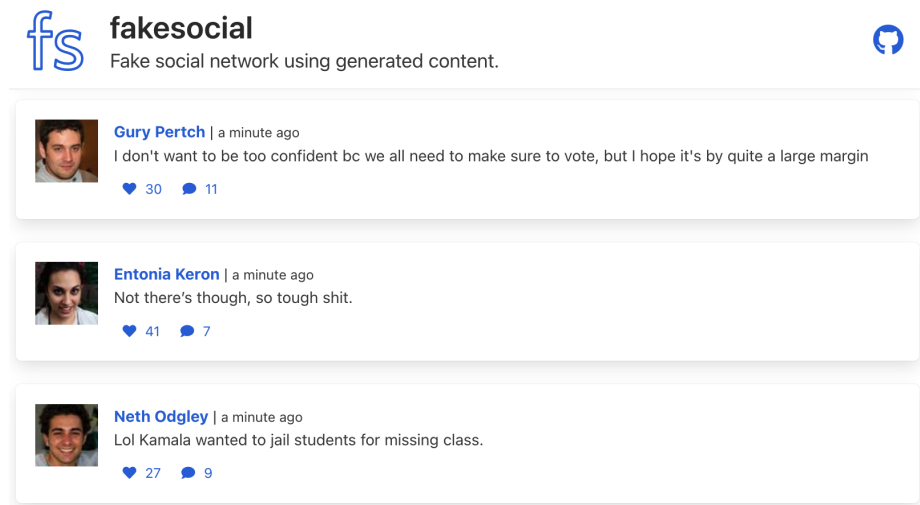


Fig. 3. Snippet of the generated social network.

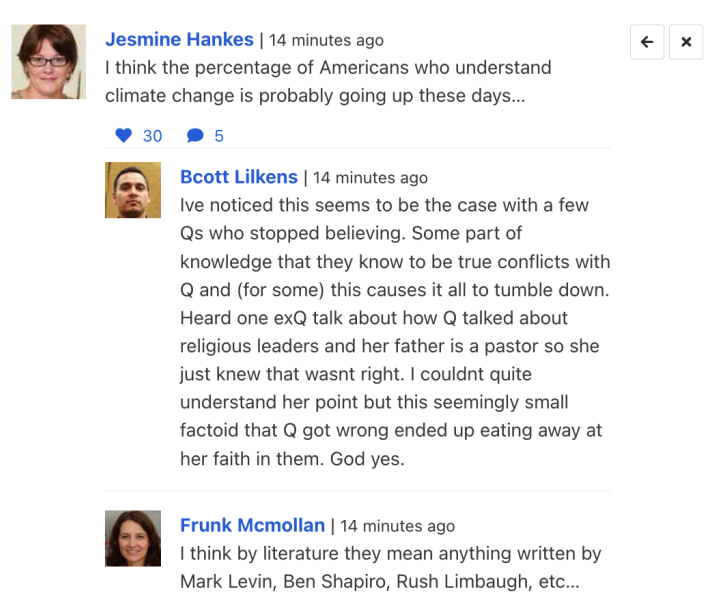


Fig. 4. Snippet of generated dialogue and network structure.

3 Results

The 3 metrics were chosen for the evaluation of the generated texts – BLEU (Post, 2018), ROUGE (Lin, 2004) and *perplexity* (Zhou and Xu, 2020). *Perplexity* is used to evaluate the language model, so this metric was applied to assess the different text generation models developed.

Table 3. BLEU and ROUGE values

Classes of text generation models	BLEU	ROUGE
Republican	0,02	0,87
Qanon	0,01	0,65
Democrat	0,03	0,48
Conservative	0,02	0,32
Mean	0,02	0,58

The BLEU estimates are shown in the 3 table. The average estimate is 0.02 and in this case, a low estimate (<10) would mean that the sample texts are quite different from the generated ones. In other words, the generated texts were similar in meaning but different in structure, and BLEU is sensitive to structural changes - for machine translation, where this metric originated, it is important that the resulting texts are as similar as possible (both in structure and in meaning) (Mathur et al., 2020). For this reason, the results obtained are weighted by the additional ROUGE metric, which usually is commonly used for the evaluation of automatic text summary generation. The ROUGE estimates are presented in table 3.

The ROUGE estimates show the percentage of sample n-grams that are also present in the generated text (e.g. 32% of the n-grams in the generated texts of the *Conservative* class are also present in the sample texts). The ROUGE results show that the generated texts can be evaluated favorably. Most of the generated texts differ in their structure (low BLEU score), but all the necessary arguments are used in the generated texts, the texts are also grammatically correct, meaningful, and correspond to the chosen ideological tendency (*Republican, Qanon, Democrat* or *Conservative*).

Table 4. Perplexity values

Classes of text generation models	Perplexity
Republican	11,37
Qanon	11,95
Democrat	11,20
Conservative	12,01
Mean	11,63

Perplexity is commonly used for model evaluation in open-domain natural language generation tasks, such as story generation and open-domain dialogue systems (Zhou and

Xu, 2020). Lower perplexity values indicate better model quality (Jang et al., 2022). In other words, the lower the perplexity value, the closer the language model is considered to be to real data. Perplexity values below 400 are considered an indication of a good quality model (Kuribayashi et al., 2021). As it can be seen in Table 4, both the perplexity values for the individual model classes and the average perplexity values are below 400, so it can be assumed that the developed text generation models are of sufficient quality. Although Fabulator at this stage is developed for the English language, it can be adapted to different languages and domains via retraining of pretrained language models.

The generated user profiles and network structures have been evaluated against the defined criteria. As it was specified in the requirements, the generated user profiles have the following attributes: *name*, *profile picture*, *workplace*, *job title*, and *place of residence*. Following the list of requirements, the types of interactions that can take place between users in a synthetic network contain *entries/messages*, *comments*, and *likes*.

As per the requirements, the synthetic graph has vertices and links between vertices. Also, each user (vertex) has between 0 and 99 "friends" and the probability is set for assigning a user's likes to each of his/her friends' entries. Finally, connections (likes, posts/comments, being a "friend") are randomly assigned to the vertices of the graph (synthetic user profiles).

4 Conclusions and future directions

There is a need for vast amounts of training and testing data for research and the development of artificial intelligence technologies. However, the usage of real-world data (especially social media data) is often protected by data protection regulations, such as GDPR. This makes a significant proportion of data unavailable to use in research. The problem can be overcome by generating synthetic data that imitates real-world data but still abides by data protection regulations. Fabulator is a synthetic generator of social media data, where synthetic graph structures and synthetic text are combined. For the text generation part, our retrained models reached 0,58 mean ROUGE value (estimation of the percentage of sample n-grams that are also present in the generated text) and 11,63 mean *perplexity* score, indicating good quality models. We got a low mean BLEU score which indicates that generated texts differ in their structure, though they are grammatically correct and meaningful. Generated graph met the requirements in terms of user profile and network structure. We believe that the combination of graph and text provides data that is more relevant and that can be used for further research of social media interactions.

Our generator initially is oriented towards Reddit data, though further research will focus on other platforms as well in order to find a more comprehensive, generalized solution. The final product should successfully generate high-quality, reliable social media interactions in the styles of different platforms to be used for the simulation of different scenarios in social media, for example, the simulation of cyber- or propaganda attacks during election periods. Such simulations can help prepare for real-life scenarios by assisting research and helping strengthen the security systems.

References

- Amarasingam, A., Argentino, M.-A. (2020). The qanon conspiracy theory: A security threat in the making, *CTC Sentinel* **13**(7), 37–44.
- Bonifati, A., Holubová, I., Prat-Pérez, A., Sakr, S. (2020). Graph generators: State of the art and open challenges, *ACM Computing Surveys (CSUR)* **53**(2), 1–30.
- Grover, A., Zweig, A., Ermon, S. (2019). Graphite: Iterative generative modeling of graphs, *International conference on machine learning*, PMLR, pp. 2434–2444.
- Iqbal, T., Qureshi, S. (2020). The survey: Text generation models in deep learning, *Journal of King Saud University-Computer and Information Sciences* .
- Jang, J., Ye, S., Lee, C., Yang, S., Shin, J., Han, J., Kim, G., Seo, M. (2022). Temporalwiki: A life-long benchmark for training and evaluating ever-evolving language models, *arXiv preprint arXiv:2204.14211* .
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. (2020). Analyzing and improving the image quality of stylegan, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119.
- Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., Inui, K. (2021). Lower perplexity is not always human-like, *arXiv preprint arXiv:2106.01229* .
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *arXiv preprint arXiv:1910.13461* .
- Li, Y., Vinyals, O., Dyer, C., Pascanu, R., Battaglia, P. (2018). Learning deep generative models of graphs, *arXiv preprint arXiv:1803.03324* .
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries, *Text summarization branches out*, pp. 74–81.
- Ma, M.-C., Lalor, J. P. (2020). An empirical analysis of human-bot interaction on reddit, *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pp. 101–106.
- Mathur, N., Baldwin, T., Cohn, T. (2020). Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics, *arXiv preprint arXiv:2006.06264* .
- Moskalenko, S., McCauley, C. (2021). Qanon, *Perspectives on Terrorism* **15**(2), 142–146.
- Nettleton, D. F., Nettleton, S. et al. (2021). Medici: A simple to use synthetic social network data generator, *International Conference on Modeling Decisions for Artificial Intelligence*, Springer, pp. 273–285.
- Post, M. (2018). A call for clarity in reporting bleu scores, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191.
- Sagduyu, Y. E., Grushin, A., Shi, Y. (2018). Synthetic social media data generation, *IEEE Transactions on Computational Social Systems* **5**(3), 605–620.
- Wu, Z., Galley, M., Brockett, C., Zhang, Y., Gao, X., Quirk, C., Koncel-Kedziorski, R., Gao, J., Hajishirzi, H., Ostendorf, M. et al. (2020). A controllable model of grounded response generation, *arXiv preprint arXiv:2005.00613* .
- You, J., Ying, R., Ren, X., Hamilton, W., Leskovec, J. (2018). Graphrnn: Generating realistic graphs with deep auto-regressive models, *International Conference on Machine Learning*, PMLR, pp. 5708–5717.
- Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation, *arXiv preprint arXiv:1911.00536* .
- Zhou, W., Xu, K. (2020). Learning to compare for better training and evaluation of open domain natural language generation models, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 9717–9724.