Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Detoxification of Russian texts based on combination of controlled generation using pretrained ruGPT3 and the Delete method

Totmina E.V.
Novosibirsk State University
Novosibirsk, Russia
e.totmina@g.nsu.ru

Abstract

This article describes our solution for the RUSSE Detoxification 2022 text automatic detoxification competition held as part of the Dialogue 2022 conference. Our approach consisted in filtering the provided training data set, fine-tuning the pretrained ruGPT3 model and selecting examples of detoxified (neutral) sentences generated with its help based on their cosine proximity and ROUGE-L to the input toxic sentence for their subsequent processing using the ruPrompts library for ruGPT-3. The final stage of processing the generated neutral comments was carried out using the Delete method - an uncontrolled detoxification model based on rules, which deleted all the remaining coarse and absentee words stored in the dictionary provided by the organizers. At the Human Evaluation stage, the system received a chrF metric value of 0.455; at the Automatic Evaluation stage - 0.505, and took eighth place at Manual Evaluation. We conducted a review and analysis of examples of detoxified sentences obtained using our model. The analysis showed that some of the generated neutral sentences in most cases lose the meaning of the original toxic sentence, and also retain either a full negative connotation or a partial one.

Keywords: sentence detoxification, ruGPT3, fine-tuning, text generation, ruPrompts, delete
DOI: 10.28995/2075-7182-2022-21-1158-1165

Детоксикация русских текстов на основе комбинации контролируемой генерации с использованием предварительно обученного ruGPT3 и метода удаления

Тотмина Е.В.
Новосибирский Государственный Университет
Новосибирск, Россия
e.totmina@g.nsu.ru

Аннотация

В этой статье описывается наше решение для конкурса автоматической детоксикации текста на русском языке the RUSSE Detoxification 2022 , проводимого в рамках конференции Dialog 2022. Наш подход заключался в фильтрации предоставленного набора обучающих данных, переподготовке предварительно обученной модели ruGPT3 и отборе примеров детоксифицированных (нейтральных) предложений, сгенерированных с ее помощью, на основе их косинусной близости и ROUGE-L к входному токсичному предложению для их последующей обработки с использованием библиотеки ruPrompts для ruGPT-3. Заключительный этап обработки сгенерированных нейтральных комментариев был проведен с использованием метода Delete - неконтролируемой модели детоксикации, основанной на правилах, которая удаляла все оставшиеся грубые и отсутствующие слова, хранящиеся в словаре, предоставленном организаторами. Анализ показал, что некоторые из сгенерированных нейтральных предложений теряют смысл исходного токсичного предложения, а также сохраняют либо полную негативную коннотацию, либо частичную.

Ключевые слова: детоксификация предложений, ruGPT3, дообучение, генерация текста, ruPrompts, delete

1 Introduction

Detoxification of a text consists in changing the content (getting rid of its rude meaning) and structure (removing obscene and rude words) of a toxic text to make it easier to read and understand, while preserving its basic idea and bringing it closer to the original meaning.

The task of *RUSSE Detoxification 2022*[1] was completed at the level of toxic comments. In this formulation, the goal is to get a neutral sentence out of a toxic one. The criteria for the complexity of the sentence include the presence of rude and toxic words that complicate the understanding of the meaning of the message, stylized graphic images to convey the emotions of the addressee (emoji), the presence of rare, ambiguous and colloquial words, the presence of anglicisms, etc.

We approach the problem in four stages. Firstly, we use the toxic and corresponding neutral comments provided by the compilers in the ternary dataset and filter them by cosine similarity and ROUGE-L (a re-oriented doubler to assess the essence for the longest common subsequence)[5, 1] metrics between toxic and neutral sentences.

We preserve pairs with high cosine similarity and average values of ROUGE L. Next, we configure the pre-trained ruGPT3 model, specifically, *sberbank ai/rugpt3medium-based-on-gpt2*[2] for the filtered dataset, similar to the setup for paraphrasing [12]. At this stage, the checkpoint-2405 model, trained on our selected data, was obtained.

Subsequent processing takes place using the *ruPrompts library for ruGPT-3*[3], in which the seed was searched by gradient descent. The model was later trained. The seed (trainable prompt) was divided into two components: the format (prompt format) and the provider (prompt provider). The model was trained using the trained ruGPT3 checkpoint-2405 model obtained at the previous stage. At the last stage of refinement, the data obtained at the previous stage were finalized using *the Delete method* proposed by the organizers as a basic one.

In the final test phase of the competition, at the Human Evaluation stage, our system received a chrF metric value of 0.455, at the Automatic Evaluation stage = 0.505, taking eighth place. In this paper we will describe our approach in more detail and analyze the quality of the generated neutral offers.

2 Related work

The detection of toxicity in user texts is an active area of research in the field of natural language processing, in particular, computational linguistics. Instagram Facebook, VK, social networks are trying to solve the problem of toxicity today. However, they usually just block such texts.

Our solution is an extension for the RuSimpleSentEval problem of simplifying sentences, the solution of which was proposed by Shatilov A. A. and Rey A. I.[10]. The approach described in the article was aimed at filtering the provided dataset, fine-tuning the pre-trained ruGPT3, modeling on it and selecting generated simple candidates based on cosine similarity and using a complex sentence as input data.

2.1 Text Style Transfer method

One of the best methods to solve the detoxification problem well is the Text Style Transfer (TST) method [6],[9]. It is worth saying that uncontrolled approaches to detoxification, taught without parallel corpora for Russian and English, already exist, as, for example, mentioned in the article[2], the authors of which collected 350 thousand offensive sentences and 7 million non-neutral sentences using a list of prohibited words. However, it is worth noting that the products of these models are often of poor quality, which in most cases does not retain significant content[6].

2.2 Deep Learning Networks for Text Generation

The first work on this topic by (dos Santos et al., 2018) is an end-to end Seq2Seq model trained on a non-parallel corpus with autoencoder loss, style classification loss and cycle-consistency loss. A more recent work by Tran et al. (2020) uses a pipeline of models: a search engine finds non-toxic sentences similar to the given toxic ones, an MLM fills the gaps that were not matched in the found sentences, and a seq2seq model edits the generated sentence to make it more fluent. Finally, Laugier et al. (2021) detoxify sentences by fine-tuning T5 as a denoising autoencoder with additional cycle-consistency loss. Dathathri et al. (2020) and Krause et al. (2020) approach a similar problem: preventing a language

[1] `https://www.dialog-21.ru/evaluation/2022/russe/`

[2] `https://github.com/sberbank-ai/ru-gpts`

[3] `https://github.com/sberbank-ai/ru-prompts`

model from generating toxic text. They do not need to preserve the meaning of the input text. However, the idea of applying a discriminator to control an LM during generation can be used for style transfer, as we show in our experiments

In tasks of summarization, similar to detoxification tasks, and headlines generation in Russian, fine-tuning of BERT-based models (BertSumAbs, mBART) is usually used [2],[4],[8]. Other description of Deep Learning methods for Text Generation tasks shown in [9] Our suggested approach aims to present a detoxified version of a user message while preserving the meaning of the original, toxic, comment.

Our proposed approach aims to present a detoxified version of a user message while preserving the meaningful significance of the original, toxic comment.

3 Task description

The task of detoxification was attributed to the task of text generation and was formulated as follows: given the texts in a toxic style, it was necessary to rephrase them into a non-toxic style, preserving the content and creating a fluent text.

3.1 Training dataset

Most text detoxification models are trained on parallel data: pairs of toxic sentences and 1-3 neutral sentences. As a training dataset, the organizers of the competition collected pairs of toxic-neutral comments. Participants were also allowed to use any additional datasets or models if they are publicly available.

3.2 Datasets for verification and testing

The data for the overall task was collected on a crowdsourcing platform. These datasets consist of pairs of one toxic sentence and one to three variants of a neutral sentence. All data is presented at the contest in github[4]. Datasets of size[5] are presented in Table 1.

Dataset type	Dataset size
Training	6,947
Validation	800
Testing	875

Table 1: Dataset size

To automatically evaluate the models, the organizers gave the following indicators:

1. **Style transfer accuracy (STA)** - the average confidence of the pre-trained BERT-based toxicity classifier for the output sentences.
2. **Meaning preservation (SIM)** - the distance of embeddings of the input and output sentences. The embeddings are generated with the LaBSE model[6].
3. **Fluency score (FL)** - the average confidence of the BERT-based fluency classifier trained to discriminate between real and corrupted sentences.
4. **Joint score (J)** - the sentence-level multiplication of the STA, SIM, and FL scores.
5. **ChrF1** metric, which is the character-level F1 score.

4 System description

4.1 Data processing and selection of candidates for model training

For our experiments, we used only Russian sentences from the training data set provided by the organizers. It contains 6,947 pairs of toxic and neutral sentences. It is worth noting that additional filtering has been applied to some of them. To select good examples, we used the following indicators to select the data of toxic and neutral proposals:

[4] https://github.com/skoltech-nlp/russe_detox_2022

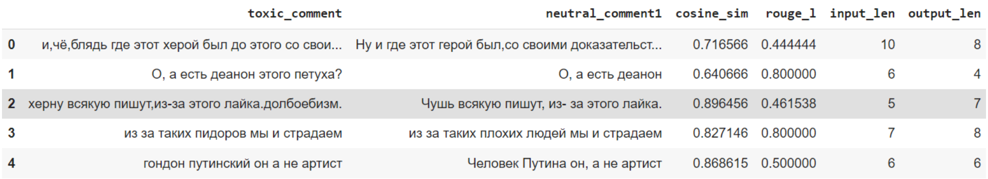
[5] The number of pairs of toxic and neutral sentences is shown for these datasets.

[6] https://huggingface.co/sentence-transformers/LaBSE

1. Cosine similarity of embeddings, obtained with BERT large model (uncased) for Sentence Embeddings in Russian language from Sberbank[7]. It shows how similar the sentences are in terms of meaning
2. ROUGE-L F1-score - Longest Common Subsequence (LCS) based statistics. It identifies longest co-occurring in sequence n-grams automatically. It shows how similar the sentences are in terms of common words.

The filtered data set was obtained by selecting pairs of sentences that make up 3869 comments (see Fig. 1) and have:

1. Cosine similarity between 0.6 and 0.99.
2. ROUGE-L between 0.1 and 0.8.
3. The length of the token of a neutral sentence, which is less than or equal to the length of the token of a toxic sentence.



	toxic_comment	neutral_comment1	cosine_sim	rouge_l	input_len	output_len
0	и,чё,блядь где этот херой был до этого со свои...	Ну и где этот герой был,со своими доказательст...	0.716566	0.444444	10	8
1	О, а есть деанон этого петуха?	О, а есть деанон	0.640666	0.800000	6	4
2	херну всякую пишут,из-за этого лайка.долбоебизм.	Чушь всякую пишут, из- за этого лайка.	0.896456	0.461538	5	7
3	из за таких пидоров мы и страдаем	из за таких плохих людей мы и страдаем	0.827146	0.800000	7	8
4	гондон путинский он а не артист	Человек Путина он, а не артист	0.868615	0.500000	6	6

Figure 1: Fragment of tabular data for the RUSSE Detoxification 2022.

4.2 A pretrained autoregressive model GPT2

We used a pretrained autoregressive GPT2-like [1] model with 350M parameters from SberDevices called rugpt3medium_based_on_gpt2 - the largest model fit into one 11GB 2080Ti GPU. This is a language model based on the transformer architecture and trained in self-supervised mode on a huge amount of text data. Compared, for example, with The T5 model [3], which uses both an encoder and a decoder, uses a 12-layer transformer architecture with only decoders.

Finetuning was done on the prepared examples from the filtered train dataset using transformers library [11]. These examples were fed into the model with the addition of special tokens (<|startoftext|> - in the beginning, <|sep|> - between toxic and detoxified sentences, <|pad|> - padding token): <|startoftext|>toxic sentence.<|sep|>detoxified sentence. After finetuning it is possible to feed into the model a prepared example as follows: <|startoftext|>New toxic sentence.<|sep|> and have the model generate a detoxified sentence. The encoded sentences were converted to a string, using a tokenizer and a dictionary with options for removing special tokens and clearing tokenization gaps.

Next, we configure the pre-trained ruGPT3 model, specifically, sberbank-ai/rugpt3medium_based_on_gp2 for the filtered dataset, similar to the setup for paraphrasing [13]. At this stage, the checkpoint-2405 model, trained on our selected data, was obtained.

Feeding the toxic sentence as a prompt into the model can generate several neutral sentences. Parameters that were used to generate candidate examples were chosen empirically and presented in Table. 2.

Parameter	Value
num_train_epochs	5
per_device_train_batch_size	4, 8, 32, 64
learning_rate	5e-5
lr_sheduleer_type	linear
warmup_steps	500

Table 2: ruGPT3 finetuning parameters

[7] https://huggingface.co/sberbank-ai/sbert_large_nlu_ru

4.3 Using the ruPrompts Library

Experiments were conducted according to data partitioning in section 4.2. Subsequent processing takes place using the ruPrompts library for ruGPT-3[8], in which the seed was searched by gradient descent, which was later trained. The seed (trainable prompt) was divided into two components: the format (prompt format) and the provider (prompt provider). The training of the model was carried out using the completed ruGPT3 model obtained at the previous stage called checkpoint-2405. The following arguments to train the model were selected. They are shown in Table. 3.

Parameter	Value
per_device_train_batch_size	2
per_device_eval_batch_size	2
eval_steps	5000
save_steps	5000
logging_steps	5000
max_steps	50000

Table 3: ruPrompts finetuning parameters

4.4 Using the Delete method

At the final stage of processing, the data obtained at the previous stage were finalized using the Delete method proposed by the organizers as the basic one. A rules-based model of uncontrolled detoxification that removes all rude and offensive words, it's already was used in such works as [7].

5 Analising the results

Code is available on GitHub[9]. The results of the data obtained at the Evaluation stage on the test dataset provided as the Evaluation result are presented in Table. 4.

Model	**ACC**	**SIM**	**FL**	**J**	**ChrF1**
e.totmina_model	0.7892	0.7298	0.7285	0.4215	55.56

Table 4: Results of the Test Evaluation stage

We also compared the results of neutralizing the toxic comments of *test.tsv* for Test Evaluation by two metrics: Cosine similarity and ROUGE-L to compare the initial results for the detoxified proposals of the organizers on training dataset and our final results (Table. 5).

Datasets	Cosine similarity	ROUGE-L
Training	0.7981	0.5355
Testing	0.8160	0.6472

Table 5: Cosine similarity and ROUGE-L comparison

Visualizations of the relations between Cosine similarity (CS) and ROUGE-L (R) together with their one-dimensional distribution separately (see Fig. 2).

[8] https://habr.com/ru/company/sberdevices/blog/596103/

[9] https://github.com/totminaekaterina/RUSSE-2022-Detoxification

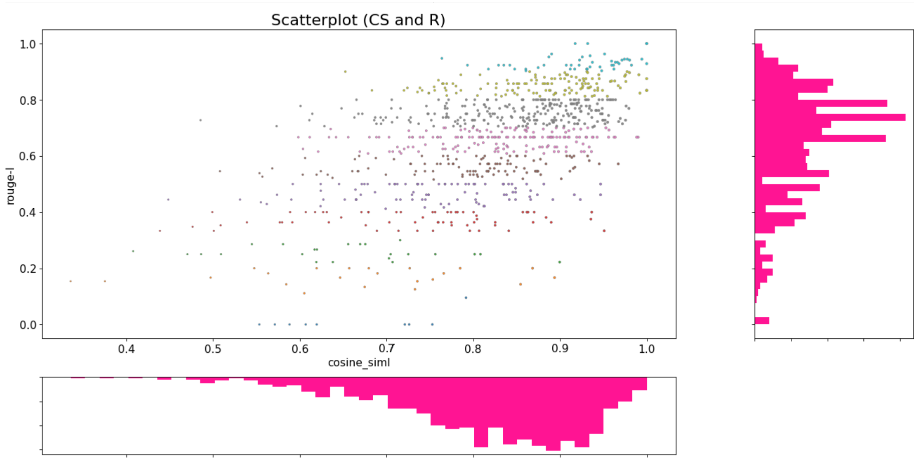


Figure 2: Visualizations of the relations between Cosine similarity (CS) and ROUGE-L (R).

We also selected some examples of the detoxified sentences predicted for test evaluation for the loss of the initial value of the input toxic sentence, as well as the complete or partial preservation of toxicity in them. The results can be divided into three groups, which are presented in Table. 6.

Result	Signs of toxicity
Complete absence of toxicity	Absence of obscene vocabulary,words with negative connotation; complete preservation of meaning
Partial loss of toxicity	Preservation of negative connotation, but with the absence of obscene vocabulary. The meaning of the sentence differs from the original, the loss of some important details, but the result can be called non-toxic
Preservation of toxicity	Complete preservation of the toxicity of the original sentence or non-sense

Table 6: Dividing the results into three groups

In table 7, you can see the proposals that completely got rid of the negative connotation and moved into the category of neutral. Table 8 shows examples of those results of proposals that were not fully detoxified or with errors, but without any serious negative features. The examples of proposals presented in table 9 can be characterized as those that have not been subjected to any restrictions at all, they have left a negative connotation and retained their original toxicity.

Original:	Пидоры когда будем дышать чистым воздухом. адмистрация это к вам
Detoxified:	Когда будем дышать чистым воздухом ,администрация это к вам
Original:	слушайте бляди если вы не вышли замуж то не хуй и пытаться
Detoxified:	слушайте если вы не вышли замуж то не нужно и пытаться

Table 7: Examples with complete absence of toxicity

Original:	Эмбрион это не человек потому что у него нет памяти, нет личности, нет эмоций, нет нихуя
Detoxified:	Ребенок это не человек потому что у него нет памяти нет личности нет эмоций нет ничего
Original:	мишустик ты бы пасть свою поганую закрыл
Detoxified:	мишустик ты бы пасть свою закрыл

Table 8: Examples with partial loss of toxicity

Original:	лживая хамелеонша судить ее надо за такое голосования
Detoxified:	Лживая хамелеонша судить ее надо за такое голосования
Original:	Послушай быдло я не смотря на свои 62 года до сих пор на летной работе врачи пока допускают
Detoxified:	Послушай быдло я не смотря на свои 62 года до сих пор на летной работе врачи пока допускают

Table 9: Examples with preservation of toxicity

6 Conclusion

In this article, we evaluated one of the approaches to detoxification of sentences by fine-tuning a pretrained ruGPT3 model and selecting generated samples based on the similarities and differences between the input toxic and output neutral sentences, using the ruPrompts library and the Delete method. At the Human Evaluation stage, the system received a chrF metric value of 0.455; at the Automatic Evaluation stage - 0.505, and took eighth place at Manual Evaluation.

Nevertheless, despite the rather high values of the indicators, the sentences created by the system completely or partially lose the original meaning of the input sentence in about half of the cases, and in most cases retain the original negative connotation. The obtained estimate may be due to the fact that the parameters we set for training models were insufficient, as well as an insufficient amount of training data set, which is confirmed by the obtained training results.

As a result, the system can be used to create examples of neutral sentences for further manual selection, but it requires some significant improvements.

References

[1] Rewon Child et al. Alec Radford, Jeff Wu. Language models are unsupervised multitask learners. 2019.

[2] Gusev Ilya Bukhtiyarov Alexey. Advances of transformer-based models for news headline generation. *arXiv*, pages 2–8, 2020.

[3] A. Roberts K. Lee S. Narang M.Matena Y. Zhou W. Li C. Raffel, N. Shazeer and J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. page 97–99, 2020.

[4] Dobrov Boris. Chernyshev Daniil. *Abstractive Summarization of Russian News Learning on Quality Media.*

[5] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. page 74–81, 2004.

[6] Rada Mihalcea Zhijing Jin Olga Vechtomova. Di Jin, Zhiting Hu. Deep learning for text style transfer: A survey. page 4–35, 2021.

[7] He He Percy Liang Juncen Li, Robin Jia. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arxiv*, page 2–12, 2018.

[8] Tutubalina Elena Malykh Valentin, Porplenko Denis. *Generating Sport Summaries: A Case Study for Russian*. 2021.

[9] Sonja Gievska Martina Toshevska. A review of text style transfer using deep learning. page 3–13, 2021.

[10] Rey A. I. Shatilov A. A. Sentence simplification with rugpt3. pages 1—-8, 2021.

[11] Victor Sanh et al. Thomas Wolf, Lysandre Debut. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Hong Kong, 2020.

[12] Andrews Martin Witteveen Sam. Paraphrasing with large language models. page 215–220, 2019.

[13] Andrews Martin Witteveen Sam. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong, 2019.