Neutralization of Evaluative Expressions Based on Dictionary Data and Distributional Models

Veronica Vybornaya
Saint Petersburg State University
Saint Petersburg
vvybornaa@gmail.com

Olga Mitrofanova
Saint Petersburg State University
Saint Petersburg
o.mitrofanova@spbu.ru

***Abstract* — Text style transfer (TST) is an important task in natural language generation, which aims to change the stylistic properties of the text while preserving the style-independent content. With the success of deep learning algorithms in the last decade, a variety of neural networks have been recently proposed for TST. If parallel data is provided, sequence-to-sequence models are usually used. However, most of the use cases do not have parallel data. Thus, this paper presents three non-parallel dataset methods for automatic identification and replacement of obscene evaluative expressions in a text, one being based on an internet dictionary Wiktionary, and two based on transformer models (BERT, GPT2). The models are then evaluated manually and automatically on a toxic dataset, extracted from a popular Russian social network VKontakte (VK). Experimental results demonstrate that the transformer-based (BERT) method has the highest average score (0.86) among style-strength and content preservation metrics.**

I. Introduction

In recent years, text style recognition, analysis and transformation has attracted not only linguists, but also many researchers in the field of computer science. In particular, the task of text style transfer (TST), which belongs to the field of natural language generation, is becoming increasingly popular.

The purpose of style transfer is to automatically control the style attributes of text while preserving the content. Language style plays a significant role within the domain of natural language processing (NLP) due to its capacity to imbue textual content with user-centric attributes. The ability to identify various language styles equips NLP models with the capacity to recognize user intentions and emotions, evaluate the formality of the text. This aptitude ultimately can improve the user contentment across diverse applications within the field.

TST has many immediate applications. For example, it can be used to make intelligent chatbots. A chatbot designed to support people can become more sympathetic and friendly if TST methods are applied. Another application is the development of writing assistants, since authors often need to edit their texts to better suit their purpose, such as making the text more professional, polite, objective, humorous, etc.

One pertinent application of this research is the addressing of offensive language, which is a prevalent issue in abusive behavior on online social media platforms. A potential solution is to provide users, who intend to post offensive messages, with a warning indicating that their content is inappropriate and will be blocked [1]. Furthermore, offering a polite alternative version of the message that can be used instead may serve as an incentive for users to reconsider their decision and refrain from using profanity.

In this study, our main objective is to develop an automated system for neutralizing offensive and inappropriate expressions in texts. We aim to provide a solution that improves readability, content preservation, and the ability to transfer text style by effectively identifying and replacing offensive language.

To achieve this objective, we will employ a combination of linguistic techniques and machine learning algorithms. The linguistic techniques will involve the analysis of dictionary entries, including the meanings and usages of words, as well as the identification of stylistic aspects such as formality, slang, or vulgarity. Machine learning algorithms will be used to apply a pretrained model that can automatically detect and neutralize profanity in texts.

The specific research questions that we address in this study are as follows:

1) *How effective is our proposed approach in automatically identifying and neutralizing profanity in texts?*

2) *What are the challenges and limitations of our approach, and how can they be addressed?*

3) *How does the neutralization process impact the readability, content preservation, and the ability to transfer text style?*

4) *What are the potential applications and real-world implications of our profanity neutralization system?*

To answer these research questions, we conduct experiments using a dataset of approximately 3000 texts collected from the VKontakte social network and assess the effectiveness of our approach. The evaluation is performed on multiple dimensions, including readability, the ability to transfer style, and content preservation. We perform both manual and automatic evaluations to comprehensively assess the effectiveness of our proposed style neutralization approach.

The rest of the paper is organized in the following way.

In Section 2, we provide a comprehensive review of the literature and discuss the current state of knowledge in the field of text detoxification.

Section 3 presents the experimental design, including details on the linguistic techniques and machine learning algorithms employed in our approach.

In Section 4, we present and discuss the results of our experiments.

Finally, in Section 5, we draw conclusions based on our findings, discuss the implications and potential applications of our research, and suggest potential avenues for future studies.

II. CURRENT RESEARCH IN THE FIELD OF STYLE TRANSFER

Existing style transfer methods are generally classified into two categories depending on the data used for training:

1) training using a parallel corpus (which contains pairs of texts of different styles with the same content);

2) training without a parallel dataset.

In this work, non-parallel TST methods are used.

Since it is difficult to collect a parallel corpus of texts and for many styles it is almost impossible to use crowdsourcing (for example, for the task of transferring a style from the language of Charles Dickens to the language of Agatha Christie), many researchers resort to using only a non-parallel mono-corpus, and apply deep learning taking into account the abovementioned limitation [2].

Existing non-parallel methods are classified depending on the type of strategy used.

The first method is explicit style-content disentanglement. In this strategy, TST models use a simple text replacement approach to generate target style texts. For example, parts of the text that are associated with the source style are first explicitly identified, and then they are replaced with new ones associated with the target style [3]. The text with the new replaced parts is then fed into the seq2seq model to generate a more natural text sequence in the target style.

An alternative method for transferring style using a single dataset is called implicit style-content disentanglement. The objective of TST models is to learn latent representations of both content and style in a given text sequence, allowing for the separation of these two elements. By combining the latent content representation from the source text with the latent representation of the desired target style, new text can be generated in the target style. Various techniques, including back translation [4], adversarial learning [5], and supervised generation, have been introduced to effectively disentangle the latent style representations.

The third strategy does not include style-content disentanglement. Recent studies have examined TST performance without disentanglement. Such techniques such as adversarial learning, reinforcement learning, probabilistic modeling and pseudo-parallel corpus construction have been applied to perform TST under this strategy [6].

As for the detoxification of texts, one of the areas of style transfer to which this work is dedicated, the first work appeared in 2018 and is an end-to-end seq2seq model trained on a non-parallel corpus [1].

Krishna et al. [7] present a methodology that involves fine-tuning pretrained language models using automatically generated paraphrase data. The STRAP (Style Transfer via Paraphrasing) models style transfer as a controlled paraphrase generation task, eliminating the need for parallel data between styles. The approach involves creating pseudo-parallel data, training style-specific inverse paraphrase models, and employing these models for style transfer.

Malmi et al. introduce MASKER, an unsupervised text-editing method for style transfer [8]. It operates without parallel source-target pairs by training masked language models for source and target domains. By identifying the most discrepant text spans in likelihood between these models, it facilitates style transformation by deleting and replacing source tokens with target masked language model. In low-resource settings, MASKER significantly improves supervised methods, enhancing accuracy by over 10 percentage points when pre-trained on MASKER-generated data. This approach offers a solution for efficient unsupervised style transfer in natural language processing applications.

Reid and Zhong [9] propose a method, which utilizes a coarse-to-fine editor with Levenshtein edit operations. The approach concurrently edits multiple spans in source text, achieving comprehensive style changes. To address the lack of parallel style text pairs, an unsupervised data synthesis procedure is introduced. Experimental results demonstrate superior performance over existing methods in sentiment and politeness transfer tasks, particularly with multi-span editing.

Regarding the progress in transferring text style for the Russian language, the first detoxification competition was initiated by Yandex in November 2021 [10]. However, the dataset provided for the competition did not include parallel data, which restricted participants from using seq2seq models. Furthermore, the evaluation system was deemed to be weak as it solely measured toxicity and similarity to the source text.

In 2022, during the “Dialogue-2022” conference on Computational Linguistics and Intellectual Technologies, a specialized track was organized that encompassed a range of practical and research tasks in NLP specifically for the Russian language. As part of this track, one of the competitions – “Detox-2022” in Russian − was focused on detoxifying texts [11]. A parallel training corpus and manual evaluation of modelswere made available for this competition.

The dataset for “Detox-2022” competition was sourced from popular Russian social networks such as Odnoklassniki, Pikabu, and Twitter. The target dataset was created through crowdsourcing, where texts were manually rewritten to remove toxic content. Four baseline options were provided, including a duplicate baseline, a rule-based approach for deletion, fine-tuning on the ruT5 model, and a continuous fast tuning approach for the ruGPT3 model. Two evaluation settings were employed in the competition: automatic evaluation based on independent metrics and a multidimensional manual evaluation.

The final phase of the competition saw the participation of ten teams. Notably, the team *gleb_shn* achieved high effectiveness in generating adversarial examples. The leaderboard revealed that the next three positions were occupied by models based on the baseline T5 system. It is noteworthy that the two models at the bottom of the leaderboard, namely the delete baseline and the model developed by the *anzak* team, demonstrated the highest content preservation. These models focused on removing or altering individual words rather than generating the output text from scratch, resulting in sentences that closely resembled the original ones. Inspired by the approach of the *anzak* team, this study employed a classifier from the detoxify library and utilized the RoBERTa-large model to substitute tokens.

III. Experimental design

To study the methods for detoxifying texts, the following steps were performed:

1) *Corpus creation:* the dataset was collected automatically using the Russian social network VK.

2) *Style transfer:* dictionary and neural methods were applied, with the former method being based on the Wiktionary online dictionary and the latter – on neural language models (BERT, GPT2).

3) *Evaluation:* the generated neutral texts were evaluated manually and automatically.

The first two methods revolve around lexical substitution, while the final is based on text-to-text generation.

In our approach, we do not include finetuning of the models, instead we conceive it as a potential avenue for future exploration in this study. However, methodologies that abstain from finetuning offer several advantages.

First, methods without finetuning are generally better at preserving the content and structure of the original text, which is a critical factor in ensuring the message or information remains intact despite stylistic modifications.

Secondly, using methods without finetuning can serve initially as a baseline to evaluate the performance and feasibility of style transfer for a particular task. It helps in understanding if finetuning is necessary.

A. Corpus Creation

To collect a corpus, it was decided to extract the required information using a parser. To meet the need, we extracted texts from two groups of the VK social network – “Палата №6”, “Подслушано” (“Ward №6”, “Overheard”), which together contain more than 160,000 user posts (over 3 500 000 tokens). The texts were extracted based on hashtags that express negative emotions, for example, “#послушано_бесит@overhear_komments” (“#overhear infuriate@overhear comments”), since such posts are more likely to contain verbal aggression. The final corpus contains 3000 texts (over 250 000 tokens).

Based on the *detoxify* multilingual library [12], 100 toxic texts were selected at random to evaluate three models.

B. Lexicon-Based Style Transfer

In this particular research study, the lexicon-based style transfer technique involved the utilization of Wiktionary, which serves as a comprehensive online dictionary powered by the wiki software. It is worth noting that Wiktionary was selected due to its versatility and multilingual capabilities.

The program itself consists of a series of steps that are sequentially executed. Firstly, the input text is tokenized and lemmatized. Subsequently, each word is checked against a stop word list in order to expedite the program's execution. If a word is not found in the stop dictionary, the corresponding word page on the Wiki dictionary is parsed. This parsing process extracts the stylistic attributes present in the dictionary entry. These attributes are then checked to ensure that they do not include any unwanted ones. Words exhibiting the following stylistic attributes, namely vulgarity, rudeness, obscenity, disparagement, and pejoration, undergo lexical substitution.

If unwanted stylistic attributes are identified, the program extracts synonyms from Wiktionary. This step is crucial as it allows for the substitution of the original word with a synonym that does not possess any undesired attributes. Finally, morphological synthesis is employed to ensure that the synonym is properly formatted to match the original word.

One prominent advantage of this approach is its rapid execution speed. On average, the program only requires approximately 15 seconds to process a text containing 100 tokens. However, there are certain disadvantages that need to be acknowledged. One notable limitation is the absence of a module for resolving lexical ambiguity of words. For instance, zoosemantic metaphors such as “овца” and “коза” (which mean “sheep” and “goat” respectively) are not recognized as undesirable words since their initial meanings denote animal names (cf. Table I). Furthermore, it is important to highlight that Wiktionary may not provide synonyms for numerous words, thereby compromising the overall quality of the model.

TABLE I. Example of output where a lexicon model fails to identify zoosemantic metaphors

Original	**Russian:** Моя подруга меня не поддерживает!Достала меня эта с*ка. Вот коза! **English:** My friend does not support me! I'm sick ofthis b*tch. She is a goat!
Generated	**Russian:** Моя подруга меня не поддерживает!Достала меня эта. Вот коза! **English:** My friend does not support me! I'm sick ofthis. She is a goat!

In the Russian language, the word “goat” among the secondary meanings has the meaning of “unpleasant girl”, in this case being a swear word.

C. BERT-Based Style Transfer

BERT (Bidirectional Encoder Representations from Transformers) is an influential family of language models that was introduced by researchers at Google in 2018 [13]. The primary objective of BERT is to facilitate the pretraining of language representations that can be effectively applied to a wide spectrum of natural language processing tasks. By leveraging the powerful Transformers architecture, BERT was trained on two tasks simultaneously: language modeling and next sentence prediction. This unique training approach empowers BERT to capture contextual information and semantic relationships within sentences in a highly efficient manner.

It is worth noting that the inspiration for this study can be traced back to the “Dialog-2022” conference [11], which hosted a competition focused on text detoxification “Detox-2022”. The present study builds upon and explores the approach put forth by the *anzak* team.

The *anzak* team solution leverages the RoBERTa-large model for its foundation. It employs a logistic regression model trained on FastText vectors from competition data to serve as a toxic word classifier. When toxic words are detected, the RoBERTa-large model is used to replace them. Replacement candidates are selected based on their cosine similarity to the toxic token. If a suitable replacement cannot be found, the toxic word is simply removed from the sentence.

Our approach distinguishes itself by utilizing the ruRoberta-large model developed by the SberDevices team [14]. This model has a dictionary size of over 50,000 words and comprises of 355 million parameters. In contrast to the *anzak* team, we do not generate multiple tokens in our methodology.

The BERT-based algorithm comprises several crucial stages. Firstly, the text undergoes tokenization, a process facilitated by the *nltk* library [15]. Subsequently, words are classified into toxic or non-toxic categories using a classifier from the detoxify library. Toxic words are then replaced with the <mask> token, utilizing the ruRoberta-large model. Finally, the algorithm generates substitutions for the masked words.

There are several advantages of the BERT-based approach. It consistently demonstrates exceptional performance in terms of the “content preservation” metric, effectively preserving the original meaning of the text. Moreover, the simplicity of the model ensures ease of implementation and usage, making it highly accessible for researchers and practitioners in the field.

However, it is important to acknowledge a significant drawback of this algorithm. Without additional training the model has the potential to generate toxic words (cf. Table II). To mitigate this risk, an effective solution would involve further training the model using a parallel corpus of texts. Additional training substantially refines the model's understanding of toxic language, ultimately minimizing the possibility of generating inappropriate or offensive content.

D. GPT-based style transfer

The field of natural language processing has witnessed remarkable progress in recent years, with the development of advanced language models. A notable example is GPT2 (Generative Pre-trained Transformer), created by OpenAI in February 2019. This powerful model exhibits a wide range of language-processing capabilities, including translation, question answering, summarization, and generating text that closely resembles human-authored content.

TABLE II. EXAMPLE OF OUTPUT WHERE A BERT MODEL FAILS TO TRANSFER STYLE

Original	**Russian:** Заткнитесь просто, ид*оты тупые, беситесвоей болтовней недалекой! **English:** Just shut up, stupid idi*ts, you infuriate mewith your small-minded chatter!
Generated	**Russian:** Ты просто , с*ка , в своей головойнедалекой ! **English:** You're just f*cking stupid in your head!

In this study, we employed a GPT2 model sourced from the Russian paraphrases library [16] to facilitate the task of paraphrasing. The model was trained using the transformers library from the SberDevices team, with a sequence length of 1024. The training process utilized 170Gb of data across 64 GPUs over a period of three weeks.

The GPT2-based style transfer methodology consists of several essential steps. Firstly, the text is tokenized using the *nltk* library. Subsequently, words are classified into toxic or non-toxic categories using a classifier provided by the detoxify library. In the case of toxic sentences, paraphrasing is performed using the GPT2 model.

While this approach offers simplicity and accessibility, caution must be exercised due to the possibility of semantic differences between the generated paraphrase and the original text (cf. Table III). Further research and refinement of the model's paraphrasing abilities.

TABLE III. EXAMPLE OF OUTPUT WHERE A GPT MODEL FAILS TO PRESERVE THE CONTEXT

Original	**Russian:** Заткнитесь просто, ид*оты тупые,бесите своей болтовней недалекой! **English:** Just shut up, stupid idi*ts, you infuriateme with your small-minded chatter!
Generated	**Russian:** Извините меня дамы и господа я насекунду. **English:** Excuse me ladies and gentlemen, Ineed to go for a minute.

IV. RESULTS

Several automatic evaluation metrics have been proposed to measure the performance of TST models. In general, these metrics evaluate models based on three criteria:

1) the ability to transfer the style of the text;

2) the amount of original non-style content that remains after applying the model;

3) language quality.

Manual and automatic evaluation were performed for each of the three models to assess their performance. In the manual evaluation, a group of 25 participants was tasked with evaluating the generated versions of the models for 5 texts based on three key criteria:

1) absence of obscene vocabulary and threats in the text;

2) the extent to which non-style-related content was preserved;

3) the naturalness of the language.

A 5-point Likert scale was used for the assessment, requiring respondents to indicate their level of agreement or disagreement with each set of statements related to the evaluated objects. This manual evaluation approach provided valuable insights into the performance of the models from a human perspective, allowing for a more comprehensive assessment of their effectiveness in achieving the desired style transfer.

The performance of the proposed models in terms of the manual metrics is shown in Table IV.

TABLE IV. THE RESULTS OF MANUAL EVALUATION

Methods	**Style transfer**	**Content preservation**	**Natural-ness of the language**	**Average score**
Wiktionary	0.46	0.82	0.5	0.59
BERT	0.58	**0.89**	**0.77**	**0.75**
GPT2	**0.59**	0.41	0.68	0.56

In this study automatic evaluation was performed as follows: the ability to transfer style was measured using a classifier from the *detoxify* library, content preservation was measured using two metrics. The language quality was not measured in the automatic evaluation, since the usual metric used (perplexity) can only be applied to pretrained models, and not on the rule-based models, since it measures the predictability of a language model based on its ability to assign probabilities to sequences of words.

The first metric for content preservation is cosine similarity. It is a metric used in text style transfer to measure the similarity between two text vectors. To use cosine similarity for text style transfer, the following steps were followed:

1) *Preprocessing:* the input and output texts were preprocessed to remove unnecessary information and convert them into vector representations. This step involved tokenization, lemmatization, and removing stopwords.

2) *Vectorization:* when the texts were preprocessed, they were converted into numerical vector representations. This step was done using a pretrained Word2Vec model from a RusVectores web-service [17], which provides pretrained distributive models. The words were therefore transformed into multi-dimensional numeric vectors.

3) *Cosine Similarity Calculation:* after the texts were converted into vector representations, the cosine similarity between the input text vector and the target output text vector was calculated. A higher cosine similarity indicates a closer similarity between the two vectors.

The second metric for content preservation is n-gram overlap. It measures the similarity between the n-grams (contiguous sequences of n words) in the input text and the generated output text. A higher n-gram overlap indicates a better preservation of content, as it suggests that the generated text retains more of the original n-grams of the text.

To use n-gram overlap for text style transfer, the following steps were followed:

1) *Preprocessing:* the input and output texts were preprocessed. The texts were tokenized, lemmatized and all the stop-words were removed.

2) *Ngram Overlap Calculation:* the n-grams in the input text were compared with the n-grams in the generated output, using the overlapy library [18].

The performance of the proposed models in terms of the automatic metrics is shown in Table V.

TABLE V. THE RESULTS OF AUTOMATIC EVALUATION

Methods	**Style transfer**	**Content preservation (word overlap)**	**Content preservation (cosine similarity)**	**Average score**
Wiktionary	0.4	0.85	0.99	0.75
BERT	0.71	**0.94**	0.94	**0.86**
GPT2	**0.79**	0.81	**0.95**	0.85

In both automatic and manual evaluations, the BERT-based method demonstrates superior performance. The disparities in the first metric between automated and manual evaluations could be attributed to the formulation of the statement, which was provided to respondents. During the manual assessment, the participants were presented with the statement: “the text comprises of profane language, insults”. Even if only one profane word was present, respondents selected “I agree”, while the classifier automatically calculated the profanity ratio, upon which binary classification occurred.

Let us delve into the findings concerning the content preservation metric. As previously mentioned, though GPT2 generates non-toxic sentences, it fails to preserve the content of the input text.

Nevertheless, according to automatic evaluation metrics, high results were obtained in preserving the content (although the sentences themselves relate to a completely different topic). This drawback became evident during manual evaluation (41% in comparison to the average 88% of automatic metrics).

Table VI presents an example from collected corpus for the evaluation to illustrate the results and purpose of the task.

TABLE VI. EXAMPLE OF OUTPUT

Original	**Russian:** Ужасно бесят все знакомые с их вопросами типа: "А что он тебе подарил?", "Почему ничего?", "И как он это объяснил?" — и т.д. **Идите в ж*пу**! Мне и так обидно до слез, а вы еще и **подъ*бываете**! **English:** I am terribly infuriated by every acquaintance asking me questions like: "What did he give you?", "Why nothing?", "And how did he explain it?" — etc. **F*ck you**! I'm already hurt to tears, and you're also **f*cking me up**!
Wiktionary output	**Russian:** Ужасно бесят все знакомые с их вопросами типа: «А что он тебе подарил? », " Почему ничего? ", " И как он это объяснил? " — и т.д . **Идите в афедрон**! Мне и так обидно до слез, а вы еще и **подмахивать**! **English:** I am terribly infuriated by every acquaintance asking me questions like: "What did he give you?", "Why nothing?", "And how did he explain it?" — etc. **Go to a lavatory**! I'm already hurt to tears, and you're also **sweeping**!
BERT output	**Russian:** Ужасно бесят все знакомые с их вопросами типа: «А что он тебе подарил? », " Почему ничего? ", " И как он это объяснил? " — и т.д . **Идите в игнор**! Мне и так обидно до слез, а вы еще и подъ*бываете! **English:** I am terribly infuriated by every acquaintance asking me questions like: "What did he give you?", "Why nothing?", "And how did he explain it?" — etc. **Go to a blacklist**! I'm already hurt to tears, and you're also f*cking me up!
GPT2 output	**Russian:** Ужасно бесят все знакомые с их вопросами типа: "А что он тебе подарил?", "Почему ничего?", "И как он это объяснил?" — и т.д. **Идите мать вашу отсюда**! Мне и так обидно до слез, а вы еще и подъ*бываете! **English:** I am terribly infuriated by every acquaintance asking me questions like: "What did he give you?", "Why nothing?", "And how did he explain it?" — etc. **Get the f*ck out of here**! I'm already hurt to tears, and you're also f*cking me up!

V. CONCLUSION

Hence, this study focused on examining approaches to neutralize evaluative expressions containing obscene words. Additionally, an experimental corpus was prepared to facilitate the use of contextualized models (RoBERTa, GPT2) and a dictionary-based approach for generating lexical substitutions. Significance of automatic and manual evaluations is determined by the evidence in favour of the approach involving the classifier from the detoxify library and the RuRoberta-large model which attained the most favorable results among the aforementioned methods. This indicates that the use of Transformers is preferable in style transfer in general and particularly in text detoxification task.

The potential domains of text detoxification models were explored, specifically chatbots and intelligent assistants for text generation. Further applications of our study deal with data filtering and refinement for training language models on speech corpora. Such models are necessary for paraphrase generation systems, for classifying emotional and neutral speech, for intent detection, etc.

The following prospects for future work in this study can been identified:

1) Enhancing BERT Training: further training of BERT on a parallel corpus can improve its language understanding capabilities and potentially lead to better performance in various natural language processing tasks.

2) Development of a Toxic Text Detoxification Web Application: creating a user-friendly web application on the Hugging Face platform that allows users to input toxic text and receive a detoxified version as an output can be a valuable tool for online communication, promoting more positive interactions.

3) Incorporating Lexical Ambiguity Module: enhancing the Wiktionary-based algorithm by incorporating a module that considers lexical ambiguity, especially in cases of zoosemantic metaphors, can lead to more accurate and context-aware language processing.

4) Exploring Other Methods from "Dialogue-2022" Conference: implementing and evaluating methods presented at the "Dialogue-2022" conference such as those utilizing the Levenshtein distance during model training.

ACKNOWLEDGMENT

The work is supported by the RSCF grant 21-78-10148 “Modelling word meaning in individual mental lexicon using distributional semantics”.

REFERENCES

[1] C. dos Santos, I. Melnyk, and I. Padhi, “Fighting offensive language on social media with unsupervised text style transfer”, in Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 189–194, 2018.

[2] Jin, Di & Jin, Zhijing & Hu, Zhiting & Vechtomova, Olga & Mihalcea, Rada, “Deep Learning for Text Style Transfer: A Survey”, Computational Linguistics”, vol. 48, pp. 1-51, 2021.

[3] J. Li, R. Jia, H. He, and P. Liang, “Delete, retrieve, generate: a simple approach to sentiment and style transfer”, in Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1865–1874, 2018.

[4] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, “Style transfer through back-translation”, in Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 866–876, 2018.

[5] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, “Style transfer in text: Exploration and evaluation”, in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[6] Hu, Zhiqiang & Lee, Roy Ka-Wei & Aggarwal, Charu & Zhang, Aston, “Text Style Transfer: A Review and Experimental Evaluation”, ACM SIGKDD Explorations Newsletter, vol. 24, pp. 14-45, 2022.

[7] K. Krishna, J. Wieting, and M. Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing

(EMNLP), P 737–762, Online, November. Association for Computational Linguistics

[8] E. Malmi, A. Severyn, and S. Rothe. 2020. Unsupervised text style transfer with padded masked language models. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), P 8671–8680, Online, November. Association for Computational Linguistics.

[9] M. Reid and V. Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. // FINDINGS

[10] Yandex CUP official website, Toxic comment classification challenge, Web: https://yandex.ru/cup/ml/analysis/#NLP.

[11] RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora, Web: https://www.dialog-21.ru/evaluation/2022/russe/.

[12] Detoxify library: Toxic Comment Classification with Pytorch Lightning and Transformers, Web: https://pypi.org/project/detoxify/.

[13] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018.

[14] RuRoberta-large model from the SberDevices, Web: https://huggingface.co/ai-forever/ruRoberta-large.

[15] Nltk library, Web: https://pypi.org/project/nltk/.

[16] Russian paraphrasers library, Web: https://pypi.org/project/russian-paraphrasers.

[17] Rusvectores Web Service, Web: https://rusvectores.org/ru/.

[18] Overlapy library, Web: https://pypi.org/project/overlapy.