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

Moscow, June 15–18, 2022

RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora

Daryna Dementieva[1,5]*, **Varvara Logacheva**[1*], **Irina Nikishina**[1]
Alena Fenogenova[3], **David Dale**[1], **Irina Krotova**[2], **Nikita Semenov**[2],
Tatiana Shavrina[3,4], **and Alexander Panchenko**[1]
[1]Skolkovo Institute of Science and Technology (Skoltech), [2]Mobile TeleSystems (MTS),
[3]SberDevices (Sber), [4]AI Research Institute (AIRI), [5]Technical University of Munich (TUM)
{daryna.dementieva, v.logacheva, irina.nikishina, d.dale, a.panchenko}@skoltech.ru,
fenogenova.a.s@sberbank.ru, i.krotova@mts.ai, nikita.semenov@mts.ru, rybolos@gmail.com

Abstract

Text detoxification is the task of rewriting a toxic text into a neutral text while preserving its original content. It has a wide range of applications, e.g. moderation of output of neural chatbots or suggesting less emotional version of posts on social networks. This paper provides a description of RUSSE-2022 competition of detoxification methods for the Russian language. This is the first competition which features (i) parallel training data and (ii) manual evaluation. We describe the setup of the competition, the solutions of the participating teams and analyse their performance. In addition to that, the large-scale evaluation allows us to analyse the performance of automatic evaluation metrics.

Keywords: Style transfer, detoxification, corpus, text generation, evaluation, competition, metrics analysis
DOI: 10.28995/2075-7182-2022-21-114-131

RUSSE-2022: первое соревнование по детоксификации русских текстов на основе параллельного корпуса

Дарина Дементьева[1,5*], **Варвара Логачева**[1*], **Ирина Никишина**[1],
Алена Феногенова[3], **Давид Дале**[1], **Ирина Кротова**[2], **Никита Семенов**[2],
Татьяна Шаврина[3,4], **и Александр Панченко**[1]
[1]Сколковский институт науки и технологий (Сколтех),
[2] Мобильные ТелеСистемы (МТС), [3]SberDevices (Сбер), [4]AI Research Institute (AIRI),
[5]Технический университет Мюнхена (ТУМ)
{daryna.dementieva, v.logacheva, irina.nikishina, d.dale, a.panchenko}@skoltech.ru,
fenogenova.a.s@sberbank.ru, i.krotova@mts.ai, nikita.semenov@mts.ru rybolos@gmail.com

Аннотация

Детоксикация текста — это задача преобразования токсичного текста в нейтральный текст с сохранением его исходного содержания. Технологии детоксификации имеют широкий спектр применений, таких как модерация вывода чатботов или перефразирование эмоционального комментария в социальной сети. Данная статья посвящена описанию соревнования моделей для детоксификации текстов RUSSE-2022. Это первое подобное соревнование, в рамках которого были доступны (1) параллельный обучающий корпус и (2) ручная оценка моделей. Мы описываем в данной работе соревнование и модели, поучаствовавшие в нем, а также анализируем их работу. Кроме того, проведенная ручная оценка качества позволила нам проанализировать автоматические метрики, использующиеся для оценки качества моделей детоксификации и переноса стиля.

Ключевые слова: Перенос стиля, детоксификация, корпус, генерация текста, оценка качества, соревнование, анализ метрик

* Equal contribution

1 Introduction

Identification of toxicity in user texts is an active area of research. Today, social networks such as Facebook[1], Instagram[2], and VK[3] are trying to address the problem of toxicity. However, they usually simply block such kinds of texts. We suggest a proactive reaction to toxicity from the user. Namely, we aim at presenting a neutral version of a user message which preserves meaningful content. We denote this task as *detoxification*.

Detoxification can be solved with Text Style Transfer (TST) (Jin et al., 2020; Hu et al., 2020) methods. This task aims at transforming the text so that its content stays the same, and its *style* (which can refer to text sentiment, author profile, degree of politeness or formality) changes. For the majority of style transfer tasks there exists no parallel data, which makes researchers train TST models on non-parallel texts (Shen et al., 2017; Wang et al., 2019; Xu et al., 2021).

Detoxification task is usually considered a variety of TST task from *toxic* to *neutral* style. There already exist unsupervised approaches to detoxification (Dementieva et al., 2021a; Dale et al., 2021) for the Russian and English languages. However, the output of these models is often of bad quality.

Russian IT company Yandex already tried to address the detoxification problem and launched the first detoxification competition. However, we extended their setup in several directions:

- We collected **a new parallel corpus** of toxic sentences and their manually written non-toxic paraphrases. This allows to solve the detoxification task using the methods developed for parallel training data (in particular, for machine translation).
- We use an **established and tested automatic evaluation setup** (Krishna et al., 2020) which agrees with the formulation of style transfer and takes into account all aspects of transfer quality. In addition to that, we use the reference-based evaluation.
- We follow the common assumption of low reliability of automatic evaluation of style transfer and make the final decision on the models quality based on the **manual evaluation**. Our work is the first attempt to use crowdsourcing for large-scale manual evaluation of a text generation model. We describe and analyse our evaluation setup.

All general information about the presented competition as well as all used code, data, and the final results can be obtained via official website.[4]

2 Related Work

RUSSE'2022 is the first competition on detoxification based on parallel corpora for Russian and has no analogies in any languages. As for the Russian, the first detoxification was launched by Yandex in november 2021 (Yandex, 2021). However, the dataset did not include parallel data that prevented participants from using seq2seq models. Moreover, their evaluation setup was weak as it only included toxicity measuring as well as similarity to the initial text and was hacked by participants.

At the same time, a lot of attempts have been made in studying toxicity for the English language. The earliest ones were several Kaggle competitions from the Jigsaw/Conversation AI team on toxicity: the "Toxic Comment Classification Challenge" (Jigsaw, 2018) in 2018, the "Unintended Bias in Toxicity Classification Challenge" (Jigsaw, 2019) in 2019 and the "Multilingual Toxic Comment Classification Challenge" (Jigsaw, 2020) in 2020. The organizers present the largest English toxicity datasets with multiple types of toxicity (toxic, obscene, threat, insult, identity hate, etc) and a multilingual test set for other languages such as Spanish, French, Italian, Russian, etc.

Since 2019 toxicity and offensive language becomes one of the central topics at SemEvals. SemEval-2019 Task 6 and SemEval-2020 Task 12 on Identifying and Categorizing Offensive Language in Social

[1] https://edition.cnn.com/2021/06/16/tech/facebook-ai-conflict-moderation-groups
[2] https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse
[3] https://vk.com/press/stickers-hate-speech
[4] https://russe.nlpub.org/2022/tox

Media (OffensEval) (Zampieri et al., 2019; Zampieri et al., 2020) attracted about 115 teams for the first year and 145 for the second year. The next SemEval competition devoted to the toxicity held in 2021: Toxic Spans Detection[5] (Pavlopoulos et al., 2021) is a task which aims at identifying the spans that make a text toxic instead of classifying whole texts in comparison to OffensEval and Jigsaw competitions. Highlighting such toxic spans can assist human moderators who often deal with lengthy comments. The 2022 year is also quite eventful for competitions on toxicity. For instance, there is a competition called "Multimedia Automatic Misogyny Identification (MAMI)" [6] which aims at identification of misogynous memes, raising a topical issue of systematic inequality and discrimination of women online. The competition examines memes as a form of hate against women, taking advantage of both text and images available as source of information.

Another SemEval 2022 task: "Patronizing and Condescending Language Detection"[7] focuses on categorizing sentences in context (paragraphs), extracted from news articles, in which one or several predefined vulnerable communities are mentioned. The task is to identify whether the unfair treatment in the media is expressed in the text and the correct category of the Patronizing and Condescending Language.

Additionally, we pay attention to the SemEval competition of this year called "iSarcasmEval: Intended Sarcasm Detection In English and Arabic" [8]. Sarcasm is omnipresent on the social web and often present in toxic texts. Determination of sarcastic texts could be also beneficial for the text detoxification process.

It can be seen that none of the previous competitions provide parallel datasets for performing detoxification and only aim at text classification and not paraphrasing. Our competition has been inspired by the Machine Translation shared tasks as it also applies parallel data and adopts some of the evaluation techniques from Machine Translation (MT) (Akhbardeh et al., 2021). It is the first parallel dataset in Russian on the topic of detoxification. In this work we present such dataset for the first time as well as the results of shared task on that data.

3 Parallel Detoxification Dataset

To perform training and automatic evaluation we provide a parallel detoxification dataset. The dataset is the core innovation of our shared task as previous detoxification shared task relied on non-aligned text corpora.

3.1 Definition of Toxicity

Our shared task deals with one particular style - toxicity. Namely, the goal is to rewrite text from toxic to neutral. What is and what is not toxic is a crucial question which shapes the training dataset and influences the performance of detoxification models. In our work we decide to consider only cases of open toxicity: open offences, use of swear and rude words. We do not focus on subtle forms of toxicity such as sarcasm or passive aggression, since they are difficult to identify not only for machines, but also for untrained human assessors. We leave work on these types of toxicity for future work.

We should warn against conflating toxicity with sentiment. Non-toxic sentences are not necessarily pleasant, they can still contain criticism such as *bad person*, *liar*, etc. Since our task is to detoxify a text while saving its content, we allow keeping negative content.

It is important to explain our understanding of toxicity to crowd workers. We use the example-based approach. Namely, instead of definitions of what is toxic we give users examples of sentences which we consider offensive and neutral. We do so in the instruction which workers need to read before doing tasks and which they can refer to later (the full text of the instruction is given in Appendix B.1). Also, since we noticed that users often skip the instruction, we ask them to take the training. It consists of examples of toxic and neutral sentences with the explanation of their label (toxic/neutral). See the examples of training questions in Appendix B.2. After that, the user passes an exam which shows if she understands the notion of toxicity correctly. We only admit users who have the result of above 80%.

[5] https://sites.google.com/view/toxicspans
[6] https://competitions.codalab.org/competitions/34175
[7] https://sites.google.com/view/pcl-detection-semeval2022
[8] https://sites.google.com/view/semeval2022-isarcasmeval

Further during the labelling, we control users by occasionally giving them control questions and reinforce their understanding of toxicity by giving training questions.

3.2 Dataset Summary

We take source (toxic) sentences for our dataset from the Russian datasets of toxic messages from various social media: Odnoklassniki (Belchikov, 2019), Pikabu (Semiletov, 2020), and Twitter (Rubtsova, 2012). The target part of the dataset are the same messages which were manually rewritten by crowd workers to eliminate toxicity.

The dataset is divided into train, development, and test sets of the following sizes:

- training: 6 948 source (toxic) sentences,
- development: 800 source (toxic) sentences,
- test: 875 source (toxic) sentences.

For each toxic sentence we have 1–3 variants of detoxification. The examples of samples collected for the task are presented in Appendix C.

3.3 Data Collection Pipeline

To collect the dataset for this competition we hired workers via Yandex.Toloka platform. We use the pipeline for the parallel detoxification data collection which was described in the work (Dementieva et al., 2021b) and tested for English. In this work we improved this pipeline and adapted it for the Russian language.

The pipeline consists of three tasks:

- **Paraphrase generation** — the workers are asked to write a neutral paraphrase of the input text. They can also select not to rewrite the input if the text is already neutral or it is difficult to extract non-toxic content. The paraphrases generated by crowd workers can be of poor quality. Therefore, we validate them using the next two tasks.
- **Content preservation check** — given two texts (the original toxic sentence and its crowdsourced paraphrase) an annotator should indicate if the content of the texts matches.
- **Toxicity classification** — given the generated paraphrase, an annotator should label it as toxic or neutral.

During the dataset collection we tried to exclude examples which are impossible to detoxify. These are (i) sentences whose meaning is offensive, (ii) sentences which aren't toxic so can't be detoxified, and (iii) sentences with unclear meaning. See the following examples:

- **Toxic content**:
  - пристрелить этих уродов без суда и следствия (*shoot these freaks without trial*)
  - а что ты с*ка умеешь, только ноги раздвигать... (*and what can you b*tch, you can only spread your legs*)
  - п*доры они в квадрате с*ки. (*f*gs are squared b*tches.*)
- **Unclear meaning**:
  - ч оз тема ч о класс ответить д лёка продаю п*зду дочери комментарий (*h oz topic h about class answer d loka sell pussy daughter comment*)

Paraphrasing sentences with toxic content cannot remove toxicity, and if we manage to remove it, the sense of such sentence will be very different from the original one.

To increase the reliability of crowdsourcing, we have each example labelled by three crowd workers. In case of paraphrase generation this gives us multiple paraphrases (some of them are filtered out later). When doing content and toxicity checks, we get multiple judgments on each example. They are further aggregated with Dawid-Skene aggregation method (Dawid and Skene, 1979) which defines the true label iteratively giving more weight to the answers of workers who agree with other workers more often. Besides the true label, this method returns the label confidence. We consider a paraphrase correct with respect to content and toxicity if it is labelled as such with the confidence of over 90%.

4 Shared Task Description

4.1 Task Formulation

Text detoxification can be considered as a kind of textual style transfer task. The style transfer task is formulated as follows. We would like to rewrite a text so that it keeps most of its content, but one particular attribute of this text (denoted as *style*) changes. The "style" can refer to various features of the text such as the level of formality, politeness, simplicity, the presence of bias or the features of the author (e.g. gender or membership in a political party). The task is usually to transfer between two "opposite" styles (toxic–neutral, formal–informal, ancient–modern), but there can exist models which support multiple exclusive or non-exclusive styles. More formally, the notion of a "style" is defined below. We deliberately, rely on a practical notion assuming that style is an automatically measurable text attribute. A more comprehensive formal definition of all various styles is a challenging task beyond the scope of our work.

Style transfer task can be formally defined as follows. We have a set of styles $S = \{s_{src}, s_{tg}\}$[9] and two collections of documents: the source corpus $D^{src} = \{d_1^{src}, ..., d_n^{src}\}$ and the target corpus $D^{tg} = \{d_1^{tg}, ..., d_m^{tg}\}$ in the styles s_{src} and s_{tg}, respectively. Let us also define the following functions. The style of a sentence is measured with $\sigma : D \to S$. A binary function $\delta : D \times D \to \{0, 1\}$ indicates the equivalence of meanings of the two styles. Finally, the function $\theta : D \to \{0, 1\}$ defines if a text belongs to well-formed sentences.

Text style transfer task is thus defined as a function $\alpha : S \times S \times D \to D$. Given a text d^{src} and its source and target styles s_{src} and s_{tg} it transforms the text to a new text d^{tg} such that:

- the style of the text is changed from the source s_{src} to the target s_{tg}: $\sigma(d^{src}) \neq \sigma(d^{tg}), \sigma(d^{tg}) = s_{tg}$,
- the contents of the original and the transformed sentences match: $\delta(d^{src}, d^{tg}) = 1$,
- the resulting sentence is well-formed (fluent): $\theta(d^{tg}) = 1$.

Therefore, a style transfer model has to optimize all three functions. Analogously, to evaluate the performance of a style transfer model, we need to check that all three conditions hold: the style is appropriately changed, the content stayed intact, and the text is fluent.

4.2 Competition Rules

The competition was opened on December 15, 2021 and lasted until February 28, 2022. It consisted of the following stages:

- **Development stage** — this stage lasted from December 15, 2021 to January 31, 2022. At this stage we made available the training and development data. The participants were invited to train their models and submit their outputs for the development set to the public leaderboard at Codalab.[10] At this stage, the models were evaluated with the automatic metrics.

[9] Style transfer task can be generalized for S with more than two styles or for continuous styles. We use the binary case for simplicity.

[10] https://codalab.lisn.upsaclay.fr/competitions/642

- **Test stage** — this stage lasted from February 1 to 14, 2022. At the beginning of this stage the participants were given access to the source part of the test set. They had two weeks to run their best-performing models on the test set and submit their answers to Codalab. The test stage leaderboard was hidden until the end of the competition.
- **Manual evaluation stage** — this stage lasted from February 14 to 28, 2022. At this stage we conducted the manual evaluation of the test answers submitted by participants and the baseline answers. The evaluation was performed via crowdsourcing. At the end of this stage we released the final leaderboard based on the results of manual evaluation.

We allowed participants to use detoxification models of any architecture. Participants were allowed to use any additional data and existing pre-trained models under open source licences.

Once results were submitted, we required participants to provide their source code and model via GitHub and also write its short description.

4.3 Baselines

We provide four baselines for detoxification task: a trivial Duplicate baseline, a rule-based Delete approach, fine-tuning on the ruT5 model and the continuous prompt tuning approach for ruGPT3 model.

Duplicate This is a trivial baseline which consists in leaving the input text intact. It provides a lower threshold for models.

Delete Delete is an unsupervised method that eliminates toxic words based on a predefined toxic words vocabulary. The idea is often used on television and other media: rude words are bleeped out or hidden with special characters (usually an asterisk). We provide both the vocabulary and the script that performs the replacement.

RuT5 Baseline Another approach is the supervised baseline based on the T5 model. We fine-tune the ruT5-base model[11] on the training part of the provided dataset.

RuPrompts This baseline is based on the ruPrompts library[12] for fast language model tuning via automatic prompt search. The Continuous Prompt Tuning method (Konodyuk and Tikhonova, 2021) consists in training embeddings corresponding to the prompts. Such approach is cheaper than classic fine-tuning of big language models. We tune the prompts for the ruGPT3-large model.[13] Pre-trained prompts for detoxification task are available online.[14]

5 Evaluation

We use two evaluation setups: automatic evaluation with reference-free and reference-based metrics and manual multi-aspect evaluation.

5.1 Automatic Evaluation

In our automatic evaluation we follow the state-of-the-art evaluation strategies. Namely, we replicate the setup of (Krishna et al., 2020). We evaluate the three parameters of style transfer quality: style of text, content preservation, and fluency of text.

Note that these three parameters exactly correspond to the TST definition components as formulated in Section 4.1: namely functions $\sigma(\cdot)$, $\delta(\cdot,\cdot)$, and $\theta(\cdot)$. The three metrics are then aggregated to a joint score. We use the following techniques.

Style ($\mathbf{STY}_a$) is evaluated with a BERT-based classifier for toxicity detection. We fine-tune the ruBERT model (Kuratov and Arkhipov, 2019) on the Odnoklassniki (Belchikov, 2019) and Pikabu (Semiletov, 2020) datasets. Style accuracy is denoted as $\sigma(\cdot)$ in Section 4.1.

[11] https://huggingface.co/sberbank-ai/ruT5-base
[12] https://sberbank-ai.github.io/ru-prompts
[13] https://github.com/sberbank-ai/ru-gpts
[14] https://huggingface.co/konodyuk/prompt_rugpt3large_detox_russe

Content ($\mathbf{SIM}_a$) is evaluated as the cosine similarity of embeddings of the source and the transformed sentences. We use embeddings generated by LaBSE model (Feng et al., 2020) because in our preliminary experiments they showed the best performance for Russian. We prefer the embedding distance over BLEU-like metrics, because (Yamshchikov et al., 2021) showed that embedding-based metrics are better correlated with human judgments than ngram-based metrics such as BLEU. We do not use references for the evaluation of content to mimic the setup where references are unavailable, which is very common for style transfer tasks. Content similarity is denoted as $\delta(\cdot, \cdot)$ in Section 4.1.

Fluency ($\mathbf{FL}_a$) Although fluency is usually evaluated as perplexity, we follow (Krishna et al., 2020) and use an acceptability classifier. In this work this classifier was trained on CoLA dataset (Warstadt et al., 2019). Since there is no such dataset for Russian, we create synthetic examples of corrupted sentences by randomly replacing, deleting or shuffling words in sentences as suggested by (Kann et al., 2018). We choose this method over perplexity, because it ranges from 0 to 1 and its greater values mean higher quality, just like metrics we use for evaluating toxicity and content. This makes combining the three metrics easier. Fluency is denoted as $\theta(\cdot)$ in Section 4.1.

Joint ($\mathbf{J}_a$) Following (Krishna et al., 2020), we combine the three metrics at the sentence level by multiplying them. Since all scores are binary, the joint score is 1 only if all three metrics are 1. Therefore, it indicates fully acceptable sentences.

$$\mathbf{J} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{STA}(x_i) \cdot \mathbf{SIM}(x_i) \cdot \mathbf{FL}(x_i) \tag{1}$$

ChrF We provide an additional reference-based metric which follows the Machine Translation evaluation setup. We choose ChrF (Popović, 2015) over BLEU, because it compares character ngrams and is more suitable for languages with rich morphology, such as Russian.

5.2 Manual Evaluation

The manual evaluation follows setups used in state-of-the-art works. We separately evaluate the three parameters of the transferred sentences, namely, their style, content, and fluency. We conduct the evaluation via crowdsourcing. For the evaluation we also use Yandex.Toloka platform.

5.2.1 Evaluation Metrics

All three parameters are evaluated at the sentence level in terms of a binary scale, where 0 refers to the bad quality in terms of the parameter and 1 is the good quality. Assessors are given the following guidelines.

Toxicity ($\mathbf{STY}_m$) The toxicity level is defined as:

- **non-toxic** (1) — the sentence does not contain any aggression or offence. However, we allow covert aggression and sarcasm. Note also that toxicity should not be mixed with the lack of formality. Even if a sentence is extremely informal, it is non-toxic unless it attacks someone.
- **toxic** (0) — the sentence contains open aggression and/or swear words (this also applies to senseless sentences).

Content ($\mathbf{SIM}_m$) In terms of content, sentences should be classified as:

- **matching** (1) — the output sentence fully preserves the content of the input sentence. Here, we allow some change of sense which is inevitable during detoxification (e.g. replacement with overly general synonyms: *idiot* becomes *person* or *individual*). It should also be noted that content and toxicity dimensions are independent, so if the output sentence is toxic, it can still be good in terms of content.

- **different** (0) — the sense of the transferred sentence is different from the input. Here, the sense should not be confused with the word overlap. The sentence is different from its original version if its main intent has changed, (cf. *I want to go out* and *I want to sleep*). The partial loss or change of sense is also considered a mismatch (cf. *I want to eat and sleep* and *I want to eat*). Finally, when the transferred sentence is senseless, it should also be considered *different*.

Fluency ($\mathbf{FL}_m$) The fluency evaluation is different from the other metrics. We evaluate it along a ternary scale with the following values:

- **fluent** (1) — sentences with no mistakes, except punctuation and capitalisation errors.
- **partially fluent** (0.5) — sentences which have orthographic and grammatical mistakes, non-standard spellings. However, the sentence should be fully intelligible.
- **non-fluent** (0) — sentences which are difficult or impossible to understand.

However, since all the input sentences are user-generated, they are not guaranteed to be fluent in terms of this scale. People often make mistakes, typos and use non-standard spelling variants. We cannot require that a detoxification model fixes them. Therefore, we consider an output of a model fluent if the model did not make less fluent than the original sentence. Thus, we evaluate both the input and the output sentences and define the final fluency score as **fluent** (1) if the fluency score of the output is greater or equal to that of the input, and **non-fluent** (0) otherwise.

Joint Score ($\mathbf{J}_m$) Finally, We aggregate the three metrics in the same Joint score as it was done for automatic evaluation.

Note that, in manual evaluation setup, we again resort to the original TST formulation based on three functions as defined in Section 4.1: $\sigma(\cdot)$, $\delta(\cdot,\cdot)$, and $\theta(\cdot)$. However, in this case, their outputs are defined not in an automatic way but rather using human judgements.

5.2.2 Crowdsourcing Setup

Each of the three parameters is evaluated in a separate crowdsourcing project. For all the projects we hire only native speakers of Russian.

Crowdsourcing tasks In the toxicity detection task (see Figure 1) we show workers the transferred sentence and ask them if it is offensive. Then, in the content similarity task we show both sentences and ask if they mean the same. Finally, we apply the fluency evaluation task to both the source and the target and compute the final fluency score from the source and target scores. While here we provide English interfaces examples, the original interfaces are presented in Appendix A.

Each sentence in each of the projects is labelled by 10 to 12 workers. We aggregate their result using Dawid-Skene aggregation method (Dawid and Skene, 1979). It takes into account the dynamically defined reliability of workers. For each example with multiple labels Dawid-Skene method returns the label and its confidence. We use only labels whose confidence is above 90%. The other labels (around 3% of all examples) are later filled by experts.

Quality Control Before admitting users to accomplishing tasks we need make sure they understand them correctly. For that purpose we devise a pipeline of training and exam tasks. First, a user needs to pass training (a set of tasks with a known label and an explanation of the task shown if the user makes a mistake) and exam (same as training, but no explanations are shown). We only admit users whose exam score is above 80%. Similarly, we control their performance with control questions during labelling. We ban users whose performance on these control question is below 70%.

Finally, we use other heuristics to control the user performance:

- **captcha** — prevents workers from using scripts and bots for labelling,
- **fast answers** — we ban users who accomplish a page of tasks in less than 15 seconds (this usually means that the user is not reading the task and is giving random answers),

- **skipped tasks** — we ban users who skip 5 or more task pages (this indicates a user who does not understand the task).

6 Participating Systems

Ten teams participated in the final phase of the competition. Here we briefly describe them. For the easier navigation in the leaderboard, we provide the models aliases which summarise the methods they use.

orzhan (ruT5-finetune) approach is based on the ruT5-base model[15]. It was fine-tuned on the part of competition train data with a learning rate 1e-5 on 15 epochs. Only the samples with fluency, similarity, and accuracy higher than 0.5 were selected from the train set. The best output is selected from 32 generated samples using beam search. It was decided not to use sampling.

NSU team (ruGPT3-filter) This team's solution uses a model based on ruGPT3. The authors filtered the dataset released by the organizers with the following heuristics: (i) cosine similarity between the original and transformed sentences ranges from 0.6 to 0.99; (ii) ROUGE-L between the sentences ranges from 0.1 to 0.8; (iii) the transformed sentence length is less or equal to the original sentence length. This dataset was used to fine-tune ruGPT3.

Mindful Squirrel (lewis) solution is based on the LEWIS framework (Reid and Zhong, 2021), a coarse-to-fine editor for style transfer that transforms text using Levenshtein edit operation. First, the sequence of coarse-grain Levenshtein edit types (keep, replace, delete or insert) was predicted for each sentence pair. Next, the resulting tags were used to train the conversational RuBERT[16] for the sequence tagging task. The ruT5-base model was trained to fill in the tokens for coarse-grain edit type *replace*.

king_menin (ruGPT3-XL) trained RuGPT3 XL[17] to generate a non-toxic text on the competition train data. The input is the concatenation of the toxic and non-toxic sentences.

[15]https://huggingface.co/sberbank-ai/ruT5-base
[16]https://huggingface.co/DeepPavlov/rubert-base-cased-conversational
[17]https://huggingface.co/sberbank-ai/rugpt3xl

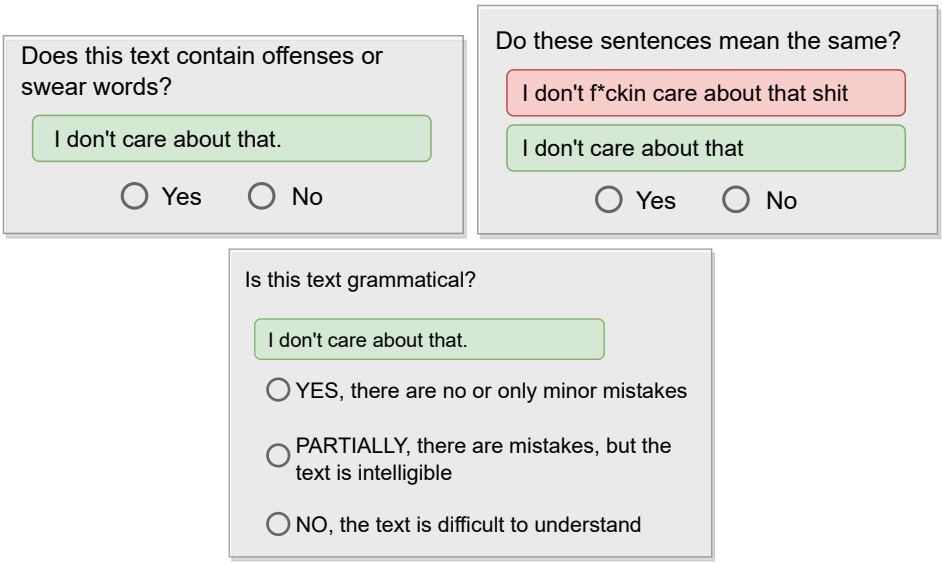


Figure 1: Design of crowdsourcing user interfaces of the toxicity detection (top left), content check (top right), and fluency check (bottom) tasks. Forms were translated from Russian to English for readability.

anzak (RoBERTa-replace) solution is based on the RoBERTa-large[18]. The logistic regression model on the FastText vectors trained on the competition data was used as a toxic words classifier. Toxic tokens were substituted by RoBERTa-large model, where the best candidates were chosen by the cosine similarity between the candidate and the toxic token. In case it was not possible to find an acceptable candidate, the toxic word was removed from the sentence.

SomethingAwful (ruT5-clean) used the ruT5-large model[19] improved by data cleaning. The preprocessing stage consitsts of emoticons and smiley filtering and removing duplicate characters. The Levenshtein Transformer (Gu et al., 2019) was used as an extra step in preprocessing to clean the ruT5-large model output.

FRC CSC RAS (ruT5-large) modified the t5 baseline. RuT5-base was replaced by ruT5-large with beam search used as inference algorithm. 20 candidates were generated for each toxic sentence, the best candidate was selected by the largest J-score metric.

barracudas (ruT5-preproc) This solution is based on ruT5-base model with additional pre- and post-processing of the texts.

gleb_shnshn (adversarial) This team devised an adversarial training setup where the training data was enriched with the artificially generated sentences which attained the highest scores of the automatic metrics.

ruPrompts-plus (ruPrompts-plus) This team advanced over the ruPrompts baseline. The solution is based on RuGPT3-XL[20] adapted to the task via prompt tuning. In particular, the participant prepended 100 and appended 20 trainable embeddings to the toxic text and passed it to the model, which was expected to output the detoxified version. These embeddings were directly optimized by gradient descent.

7 Results

The primary goal of our competition is to evaluate the models and understand which approach is more promising. Here we compare the performance of models in terms of manual and automatic metrics. Besides that, since we have both manual and automatic scores, we evaluate the performance of metrics themselves.

7.1 Models Performance

Table 1 shows the performance of the participating models and our baselines in terms of the automatic metrics. The adversarial example generation (**gleb_shnshn**) turns out to be very effective — it attains the highest scores of all metrics, thus yielding the highest J_a score. The next three places in the leaderboard are taken by the models based on our baseline ruT5 system (**orzhan**, **FRC CSC RAS**, and **SomethingAwful**). This suggests that this model is very efficient. Notice that the human references are below the majority of models in terms of all metrics except ChrF whose score for the human references is the highest by a large margin.

It is also important to note that the highest content preservation is demonstrated by two models from the bottom of the leaderboard, namely, the **Delete** baseline and **anzak** team's model. Both of them do not generate the output text from scratch but only remove or change individual words. This approach yields sentences which are very similar to the original ones.

The manual scores (see Table 2) provide a completely different result. There, the human references are significantly better than other models, but closely followed by the solution by the **SomethingAwful** team. This team is the only team whose solution succeeded in outperforming the ruT5-based baseline model. The winning team's model is also based on ruT5 (although they use ruT5-large), but with the additional preprocessing. The model of the **FRC CSC RAS** team, which got the 3rd best result in terms of fluency

[18] https://huggingface.co/sberbank-ai/ruRoberta-large
[19] https://huggingface.co/sberbank-ai/ruT5-large
[20] https://huggingface.co/sberbank-ai/rugpt3xl

Team	Method	ACC_a	SIM_a	FL_a	J_a	ChrF
gleb_shnshn	adversarial	**0.97**	**0.94**	**0.96**	**0.87**	0.53
orzhan	ruT5-finetune	**0.98**	0.86	**0.97**	0.82	0.55
FRC CSC RAS	ruT5-large	**0.95**	0.86	**0.97**	0.78	0.57
SomethingAwful	ruT5-clean	**0.95**	0.82	0.91	0.71	0.57
Mindful Squirrel	lewis	0.93	0.80	0.88	0.66	0.56
king_menin	ruGPT3-XL	0.94	0.73	0.89	0.61	0.50
baseline	RuT5	0.80	0.83	0.84	0.56	0.57
ruPrompts-plus	ruGPT-XL+ruprompts	0.80	0.80	0.83	0.54	0.56
baseline	ruPrompts	0.81	0.79	0.80	0.53	0.55
barracudas	ruT5-preproc	0.85	0.76	0.78	0.52	0.53
human references	manual annotation	0.85	0.72	0.78	0.49	**0.77**
NSU team	ruGPT3-filter	0.83	0.76	0.76	0.48	0.51
anzak	RoBERTa-replace	0.57	**0.89**	0.91	0.44	0.54
baseline	Delete	0.56	**0.89**	0.85	0.41	0.53
baseline	Duplicate	0.24	1.00	1.00	0.24	0.56

Table 1: The performance of the participating models in terms of automatic metrics, sorted by J_a metric. The values **in bold** show the highest value of the metric with the significance level of $\alpha = 0.05$.

and content preservation, is also based on ruT5-large model. This confirms that large pretrained models with fine-tuning on parallel data are a very strong baseline which is hard to beat.

Interestingly, the **adversarial** model whose automatic scores are the highest, in fact produces sentences of a very low quality. This shows that automatic metrics can be "fooled" and should not be used as an ultimate evaluation technique.

In terms of the quality of style change, the model of the **Mindful Squirrel** team yielded the best result which was only outperformed by human references. This model uses a word classifier which decides if a word should be changed or left intact during style transfer. This allows to focus on toxic words.

Overall, the evaluation shows that the models based on **ruT5** fine-tuned on parallel data are the most successful. The two teams that used **ruGPT3** could not approach the results of the competitors. The tuning of prompts is still less efficient than tuning of models. The models based on explicit edit operations are only moderately successful.

7.2 Automatic vs Manual Metrics

The automatic and manual metrics (Tables 1 and 2) provide very diverse results. This suggests that they are weakly correlated.

We check this assumption by computing the Spearman ρ correlations for document-level scores of all metrics. We put in bold all high correlations (p-value ≤ 0.05) in Table 3. We clearly see that none of automatic metrics correlate with their manually measured counterparts. On the other hand, manual style and content metrics are correlated with ChrF score. This suggests that ChrF can be used as an automatic evaluation score. On the other hand, ChrF is not sensitive to sentence style, which means that it can be deceived (for example, the trivial Duplicate baseline performs on par with strong T5-based models in terms of ChrF). However, the power of ChrF was also claimed by (Briakou et al., 2021).

The sentence-level correlations show a slightly different picture. The highest correlation is seen for the style metric, the Spearman ρ score of automatic and manual judgments is 0.418 (moderate correlation). The manual and automatic sentence-level similarity, fluency, and joint scores show very weak or no correlation: 0.251, 0.015, and 0.141, respectively.

However, sentence-level correlations between corresponding manual and automatic metrics differ significantly across models (see Figure 2). We see that automatic and manual toxicity scores are much

Team	Method	ACC_m	SIM_m	FL_m	J_m
human references	manual annotation	**0.89**	0.82	**0.89**	**0.65**
SomethingAwful	ruT5-clean	0.79	**0.87**	**0.90**	0.63
baseline	RuT5	0.79	0.82	**0.92**	0.61
FRC CSC RAS	ruT5-large	0.73	**0.87**	**0.92**	0.60
Mindful Squirrel	lewis	**0.82**	0.79	0.85	0.58
ruPrompts-plus	ruGPT-XL+ruprompts	0.78	0.81	**0.90**	0.57
orzhan	ruT5-finetune	0.80	0.78	0.87	0.56
barracudas	ruT5-preproc	0.79	0.72	0.78	0.51
king_menin	ruGPT3-XL	0.81	0.70	0.90	0.50
baseline	ruPrompts	0.80	0.70	0.87	0.49
NSU team	ruGPT3-filter	0.77	0.72	0.83	0.45
anzak	RoBERTa-replace	0.43	0.62	0.79	0.17
baseline	Delete	0.39	0.71	0.73	0.16
baseline	Duplicate	0.11	1.00	1.00	0.11
gleb_shnshn	adversarial	0.25	0.13	0.24	0.02

Table 2: Manual evaluation of the participating models, the models are sorted by the J_m metric. The values **in bold** show the highest value of the metric with the significance level of $\alpha = 0.05$.

Metric	STA_a	SIM_a	FL_a	J_a	ChrF
STA_m	0.376	**-0.776**	-0.398	0.278	0.223
SIM_m	-0.046	0.031	0.190	0.000	**0.789**
FL_m	-0.083	-0.032	0.288	0.070	**0.619**
J_m	0.326	-0.495	-0.211	0.350	**0.735**

Table 3: Spearman's correlation coefficient between automatic VS manual metrics on system level. Bold numbers denote the statistically significant correlation (p-value ≤ 0.05).

better correlated for the **Delete** and **anzak** models, which are the only models to explicitly remove or replace toxic words identified by a classifier or via a manually compiled list of toxic words. These models apparently produce texts which are easy to classify correctly. Conversely, **gleb_shnshn** model and **human references** are the most difficult to classify. The former deliberately "fools" the classifier with artificial examples, while the latter contains non-trivial phrases whose level of toxicity is difficult to grasp automatically.

Analogously, the similarity scores are also better correlated for the **anzak** model which leaves the majority of words intact, so for it similarity boils down to word matching. On the other hand, T5-based models produce non-trivial paraphrases. These T5 outputs are also difficult to correctly classify for fluency, unlike the models based on word replacements (**anzak** and **Delete**). Overall, we see that it is more difficult to correctly classify outputs of *better-performing models* and *models based on large pre-trained language models* than the simple baseline approaches. This suggests that the automatic evaluation might fail exactly where we need it most, i.e. in discriminating between the good models.

8 Conclusions

We organised a competition on text detoxification for the Russian language. To the best of our knowledge, this is the second such competition. This is also the first detoxification challenge that used manual evaluation. For the needs of competition we created the first parallel Russian corpus for detoxification enabling the use of supervised machine translation approaches to this task.

Our analysis of model performances showed that the best result is attained by models based on the pre-trained ruT5 model fine-tuned on our parallel data. This model produces sentences which were evaluated closely to the human references. This shows that pre-trained Transformers are very powerful and are

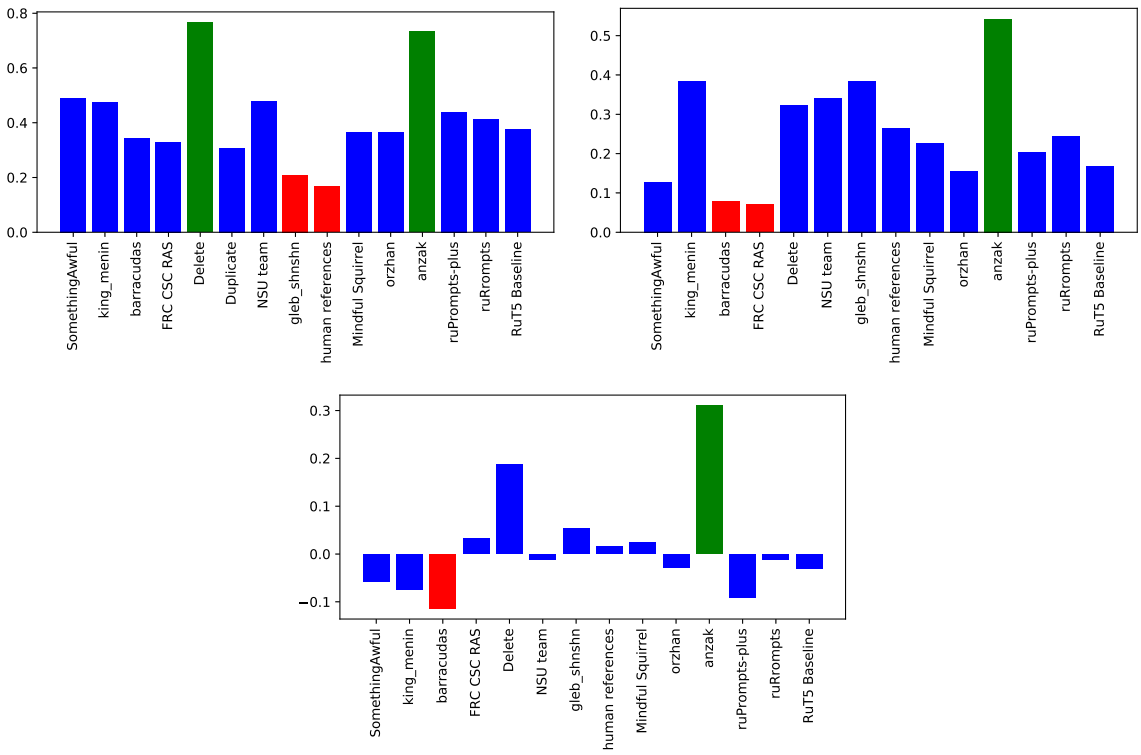


Figure 2: Correlations between automatic and manual metrics at the sentence level for different models: correlation of style accuracy scores (top left), correlation of similarity scores (top right), correlation of fluency scores (bottom). Red and green bars indicate the lowest and the highest values, respectively.

difficult to beat.

We conducted an evaluation of detoxification models for Russian using both automatic and manual metrics. This allowed us to analyse the relationship between the metrics and assess the suitability of automatic metrics for evaluation.

Our analysis shows that the metrics are overall weakly correlated with the human judgements both at the system and the sentence level. We found that ChrF score has a strong correlation with the joint score of style, content, and fluency. Thus, ChrF could be used as a proxy for manual evaluation, but its lack of correlation with the style score makes this metric vulnerable to attacks. We also discovered that the correlation of manual and automatic scores varies for different models. This shows the necessity to consider diverse style transfer models for metrics analysis.

Overall, although the state-of-the-art evaluation setup for style transfer (three parameters and the joint score combined from them) is conceptually correct, the current performance of automatic metrics is insufficient to use it as a replacement for manual evaluation. More research is needed to better fit the quality of manual evaluation.

Acknowledgements

This work was supported by the joint MTS-Skoltech laboratory on AI. The manual evaluation was supported by a Yandex.Toloka research grant.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). // *Proceedings of the Sixth Conference on Machine Translation*, P 1–88, Online, November. Association for Computational Linguistics.

Anatoly Belchikov. 2019. Russian language toxic comments. https://www.kaggle.com/blackmoon/russian-language-toxic-comments. Accessed: 2021-07-22.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 1321–1336, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 7979–7996, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

A. P. Dawid and A. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021a. Methods for detoxification of texts for the russian language. *Multimodal Technologies and Interaction*, 5(9).

Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021b. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. // *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. *Advances in Neural Information Processing Systems*, 32.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2020. Text style transfer: A review and experimental evaluation. *arXiv preprint arXiv:2010.12742*.

Jigsaw. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge. Accessed: 2022-03-16.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification. Accessed: 2022-03-16.

Jigsaw. 2020. Jigsaw multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification. Accessed: 2022-03-16.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! // *Proceedings of the 22nd Conference on Computational Natural Language Learning*, P 313–323, Brussels, Belgium, October. Association for Computational Linguistics.

Nikita Konodyuk and Maria Tikhonova. 2021. Continuous prompt tuning for russian: how to learn prompts efficiently with rugpt3? // *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts*.

Kalpesh Krishna, John Wieting, and Mohit 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.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. // *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, P 59–69, Online, August. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. // *Proceedings of the Tenth Workshop on Statistical Machine Translation*, P 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Machel Reid and Victor Zhong. 2021. LEWIS: Levenshtein editing for unsupervised text style transfer. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 3932–3944, Online, August. Association for Computational Linguistics.

Yulia Rubtsova. 2012. Avtomaticheskoe postroenie i analiz korpusa korotkih tekstov (postov mikroblogov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora. // *Inzheneriya znaniy i tekhnologii semanticheskogo veba, T.1*, P 109–116.

Aleksandr Semiletov. 2020. Toxic russian comments. https://www.kaggle.com/alexandersemiletov/toxic-russian-comments. Accessed: 2021-07-22.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Haoran Xu, Sixing Lu, Zhongkai Sun, Chengyuan Ma, and Chenlei Guo. 2021. Vae based text style transfer with pivot words enhancement learning.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220, May.

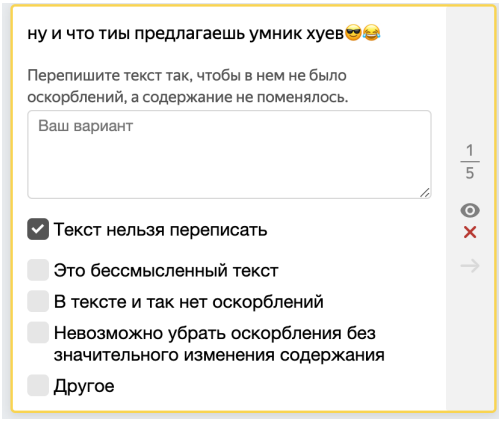
Yandex. 2021. Toxic comment classification challenge. https://yandex.ru/cup/ml/analysis/#NLP. Accessed: 2022-03-16.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). // *Proceedings of the 13th International Workshop on Semantic Evaluation*, P 75–86.

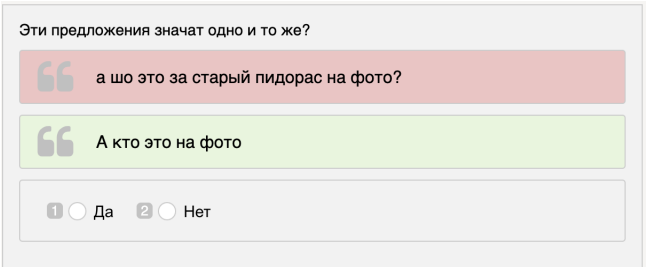
Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). // *Proceedings of SemEval*.

A Labeling Pipeline Instructions

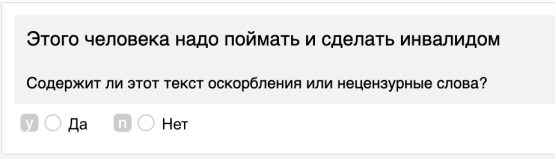
This appendix contains the illustration of all labeling tasks at Yandex Toloka platform in original Russian language: (i) detoxificated paraphrase generation (Figure 3a); (ii) content preservation check (Figure 3b); (iii) toxicity classification (Figure 3c); (iv) fluency check (Figure 3d).



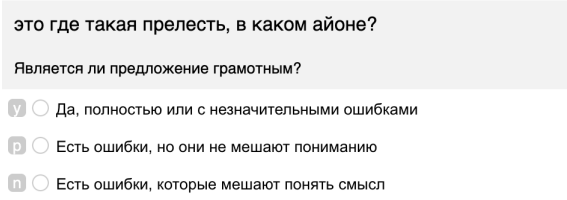
(a) Generation of paraphrases.



(b) Content preservation check.



(c) Toxicity classification.



(d) Fluency check.

Figure 3: Original Russian interfaces in Yandex Toloka platform for labeling.

B Toxicity Labelling

This appendix contains the additional information on our ways of delivering the notion of toxicity to crowd workers. In section B.1, we provide the full instruction for the toxicity labelling task which was used to label the sentences when collecting the training data and also when evaluating the outputs of models participating in the competition. Section B.2 contains the examples which were used for user training.

B.1 Labelling Instructions

You are asked to read sentences and define if they contain offences, swear or rude words.

Attention! Non-offensive sentences can contain criticism and have a negative sentiment.

Examples

Offensive sentences:

- Интернетная шпана, не тебе меня поучать. (*You internet punks, it's not for you to lecture me*)
- Какая подписка, что ты несёшь, поехавший? (*What kind of subscription, what are you talking about, id**t?*)

- Как минимум два малолетних дегенерата в треде, мда. (*At least two juvenile degenerates in a thread, hmm.*)
- Вообще не вижу смысла спорить по поводу дюймов, х**вый там ips или нет, машина не этим интересна. (*In general, I see no reason to argue about inches, no matter if ips is sh*t or not, it's not the main thing about a car.*)

Neutral (inoffensive) sentences:

- У нас в есть убунты и текникал превью. (*We have ubuntu and technical previews.*)
- он тоже был храбрым! (*he was brave too!*)
- Это бессмысленно, ты же знаешь (*It's pointless, you know*)
- Если он мечтает съесть своих соседей то это плохо. (*If he dreams of eating his neighbors, then this is bad.*)

B.2 Training Examples

- а так реагировать на простое фото это п*здец, по моему фляга у тебя свистит п*здец как (*And such reaction to a simple photo is f*cked up, in my opinion you're retarded*)
  - Label: toxic
  - Hint: Use of obscenity (*f*cked up*) and insult (*retarded*)
- тварина *баная если это ее слова (*F*cking beast if that's her words*)
  - Label: toxic
  - Hint: Use of insult (*f*cking beast*)
- ну ты глупый (*You're stupid*)
  - Label: toxic
  - Hint: Insult of the interlocutor (*stupid*)
- Лжецы. Вот и что с такими делать? (*Liars. So what should we do to them?*)
  - Label: non-toxic
  - Hint: *Liars* is not an insult, but rather a statement of fact and not a swear word.
- очень плохой человек тот кто со своим ружьем такую птицу убивает (*A very bad person is the one who kills such a bird with his gun*)
  - Label: non-toxic
  - Hint: *Bad person* is not an insult.
- Странная, люди просто салатом делятся (*You are strange, people are just sharing their salads*)
  - Label: non-toxic
  - Hint: No offences or swear words.

C Russian Detoxification Parallel Corpus Samples

This appendix presents several samples from collected corpus for the competition to illustrate the nature of the data and purpose of the task.

Original	из-за таких п*доров мы и страдаем *(we suffer because of such f*ggots)*
Paraphrases	из-за таких плохих людей мы и страдаем *(we suffer because of such bad people)* Из-за таких людей мы и страдаем *(we suffer because of such people)* из за таких как он мы и страдаем *(we suffer because of people like him)*
Original	лучше б руководство выкинули бы на хр*н!!!! *(it would be better if the leaders were thrown out to h*ll!!!)*
Paraphrases	лучше б руководство выкинули бы *(it would be better if the leaders were thrown out)* Лучше бы руководство убрали *(it would be better if the leaders were removed)* Лучше б руководство поменяли бы!!! *(it would be better if the leaders were changed!!!)*
Original	да вообще пох*й сколько этот д*лбоеб получает *(I don't f*cking care how much this motherf*cker gets paid)*
Paraphrases	да вообще все равно сколько он получает *(I don't much care how much he gets paid)* Мне всё равно сколько он получает. *(I don't care how much he gets paid)*
Original	бл*ть, развели срачь из них*я. *(d*mn, they made a f*cking quarrel out of nothing)*
Paraphrases	Развели ссору из ничего. *(they made a quarrel out of nothing)* Развели ругань из нечего *(they made a fight out of nothing)*
Original	ты мужик ох*ел что пишешь *(dude, what the f*ck are you writing)*
Paraphrases	Зачем ты такое пишешь *(why are you writing this)* Ты мужчина обалдел что пишешь *(man, are you nuts, what are you writing)*
Original	что за п*дик в судейской коллегии??? *(what a f*ggot is in the judiciary???)*
Paraphrases	кто в судейской коллегии??? *(who is in the judiciary???)* Что за человек в судейской коллегии? *(what a man is in the judiciary?)*
Original	дэтих св*лочей надо расстрелять за жестокое обращение с животным! *(these b*stards should be shot for cruelty to animals)*
Paraphrases	Этих людей нужно наказывать за жестокое обращение с животными! *(these people should be punished for cruelty to animals)*
Original	на х*я такое выкладывать, зто и дети будут смотреть д*лбоебы *(what the h*ll do you need to post this, the kids will watch it, motherf*ckers)*
Paraphrases	Зачем такое выкладывать, это и дети будут смотреть *(Why do you need to post this, the kids will watch it)*

Table 4: Examples of detoxified sentences from the collected parallel corpus.