A Study on Manual and Automatic Evaluation for Text Style Transfer: The Case of Detoxification

Varvara Logacheva[1*], Daryna Dementieva[1,5*], Irina Krotova[2], Alena Fenogenova[3], Irina Nikishina[1], 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)
{v.logacheva, daryna.dementieva, irina.nikishina, a.panchenko}@skoltech.ru,
fenogenova.a.s@sberbank.ru, rybolos@gmail.com, i.krotova@mts.ai

Abstract

It is often difficult to reliably evaluate models which generate text. Among them, text style transfer is a particularly difficult to evaluate, because its success depends on a number of parameters. We conduct an evaluation of a large number of models on a detoxification task. We explore the relations between the manual and automatic metrics and find that there is only weak correlation between them, which is dependent on the type of model which generated text. Automatic metrics tend to be less reliable for better-performing models. However, our findings suggest that, ChrF and BertScore metrics can be used as a proxy for human evaluation of text detoxification to some extent.

1 Introduction

There exist many Natural Language Processing (NLP) tasks whose output is a text (dialogue, summarization, etc.). They often adopt the evaluation techniques from Machine Translation (MT). Namely, researchers often compare the output of a model with a pre-defined reference answer and measure the model quality as the similarity to this reference. The similarity can be computed at the level of words and phrases (e.g. BLEU or METEOR) or be more semantically motivated and compare the embeddings (e.g. BertScore or BLEURT).

This approach has a number of drawbacks which make it inapplicable to some generation tasks, e.g. style transfer. This is a task of changing a text such that its meaning stays the same and the *style* changes. Style can refer to any attribute concerning only the form of the text (e.g. degree of formality or politeness) or its content (e.g. sentiment, author features, etc.). When evaluating the output of a style transfer model, we need to pay attention to both the style change and the content preservation. The traditional MT evaluation metrics mainly check the semantic similarity, which makes them unsuitable for style transfer.

There exist evaluation metrics (Krishna et al., 2020) which were devised to consider all important aspects of style transfer quality (style, semantic similarity and sometimes fluency). However, they heavily rely on automatic models (e.g. style classifier) whose performance is not perfect. Many works acknowledge the low reliability of such metrics and arrange manual evaluation to get the objective information on the models performance. Unfortunately, such evaluation is laborious and cannot be conducted often, so during development of models researchers still have to resort to automatic metrics.

Although works on style transfer acknowledge that automatic evaluation metrics are unreliable, there is little work on the analysis of their performance. There exist analysis of content preservation metrics (Yamshchikov et al., 2021) and of all style transfer evaluation metrics (Briakou et al., 2021a). The latter work provides an evaluation where metrics are tested on different systems and different style transfer directions.

We further extend this line of research by testing the evaluation metrics on a new style transfer task (detoxification) and a new language (Russian). For this comparison we create a large parallel corpus for detoxification. We compare the performance of models based on different principles, which allows more robust evaluation. Furthermore, since we compare a large number of models, we can understand to what extent the automatic metrics can *rank* the models correctly. Besides that, due to the large number of tested models we decided to use crowdsourced evaluation instead of experts. We describe our crowdsourcing annotation setup and analyse the performance of crowd workers. Finally, the large-scale evaluation allows us to gain insights on the performance of various style transfer models. The research was based on the data of a competition of detoxification models for the Russian language

[*] Equal contribution

Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 90 - 101
May 27, 2022 ©2022 Association for Computational Linguistics

organized by the authors of this paper.[1]

2 Evaluation

2.1 Style Transfer Formulation

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 (polite–impolite, positive–negative), but there can exist models which support multiple exclusive or non-exclusive styles.

Style transfer task can be formally defined as follows. We have a set of styles $S = \{s_{src}, s_{tg}\}$[2] and two corpora $D^{src} = \{d_1^{src}, ..., d_n^{src}\}$ and $D^{tg} = \{d_1^{tg}, ..., d_m^{tg}\}$ in the styles s_{src} and s_{tg}, respectively. Let us also define the following functions. The style of a sentence is measured with $\sigma : D \rightarrow S$. A binary function $\delta : D \times D \rightarrow \{0, 1\}$ indicates the equivalence of meanings of the two styles. Finally, the function $\theta : D \rightarrow \{0, 1\}$ defines if a text belongs to well-formed sentences.

Text style transfer task is thus defined as a function $\alpha : S \times S \times D \rightarrow D$. 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. However, these three conditions are often inversely correlated (Pang and Gimpel, 2019). This makes style transfer evaluation a notoriously difficult problem. Since the three conditions have to be explicitly checked, we cannot adopt the techniques used for the evaluation of other text generation models. In this work we make all evaluation on a detoxification task for which more broad definition of style transfer is fully applicable.

2.2 Automatic Evaluation of Style Transfer

In earlier works, reference-based evaluation metrics were considered a holistic evaluation technique (Li et al., 2018), by analogy with Machine Translation. Even some recent works (Sudhakar et al., 2019; Zhu et al., 2021) use BLEU or other metrics such as GLEU as the only means of evaluation. Unfortunately, they often cannot control style. Thus, it became obvious that both content and style have to be directly evaluated.

Some works settle for mere evaluation of style and content (Malmi et al., 2020; Zhang et al., 2020b). However, more often these two metrics are combined by computing their geometric or harmonic mean, as first suggested by (Xu et al., 2018). This technique is often used to get the joint quality score (Riley et al., 2021; Huang et al., 2021; Lai et al., 2021a,b). Many (although not all) works also evaluate the fluency of the generated text. This is almost exclusively done via computing perplexity of text in terms of a language model (e.g. GPT-2 (Radford et al., 2019)). The only alternative used in style transfer works is the use of classifier of linguistic acceptability (Krishna et al., 2020) trained on CoLA dataset (Warstadt et al., 2018). Fluency is sometimes also included to the joint score together with the style and content preservation. (Pang and Gimpel, 2019) compute it as a document-level geometric mean, and (Krishna et al., 2020) multiply the sentence-level scores. In our work we use the latter approach.

2.3 Manual Evaluation of Style Transfer

The researchers have come to a conclusion that these automatic metrics cannot provide an objective evaluation. It has become a de-facto standard to enhance the automatic evaluation with the human evaluation experiments.

There are two main human evaluation scenarios. Outputs of two models can be evaluated side by side, in this case the authors report the number of wins of each of the models (i.e. the number of cases where a particular model generated a better text) and the number of ties (Zhu et al., 2021; Li et al., 2019; Cheng et al., 2020). Alternatively, the outputs of different models are evaluated in-

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

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

dependently. In this case the assessors evaluate the outputs along three parameters: style, content preservation, and fluency. The parameters are often evaluated in terms of a 1-to-5 scale (Zhou et al., 2020; Madaan et al., 2020; John et al., 2019; Lee et al., 2021; Ma et al., 2021). Sometimes the style is evaluated in terms of a 7-value scale (from -3 to 3), content preservation takes values from 1 to 6 (Chawla and Yang, 2020; Briakou et al., 2021b). Other scales are also possible. Besides that, the three individual metrics can be evaluated using the side-by-side scenario (Sudhakar et al., 2019; Lin et al., 2020).

3 Detoxification Competition Details

The evaluation was conducted under the scope of a competition of detoxification models for the Russian language (Dementieva et al., 2022).[3] For this competition we created a Russian parallel corpus of toxic sentences and their manually written non-toxic equivalents. We also developed several baselines.

3.1 Parallel Dataset

We collected a parallel Russian dataset for detoxification for this competition. The corpus was collected via the Yandex.Toloka[4] crowdsourcing platform. We used the data collection setup described by (Logacheva et al., 2022) There, the crowd workers were asked to rewrite a sentence so that it preserves its content, but does not sound toxic. Then other crowd workers checked the rewritten sentence for toxicity and semantic similarity with the original one. The platform of Yandex.Toloka has a special mark for cases of inappropriate and toxic content. Thus, all the crowd workers were notified about possible unethical context of the task and we get approvals for the experiment.

As it was noted, we need the toxic and corresponding neutral sentences to be semantically similar. Therefore, during the generation of the dataset we ask crowd workers to rewrite the sentence in a non-offensive way and keep its content. If it is impossible to detoxify a sentence, a worker can choose to not change it. Such sentences are not included to the resulting dataset. All the generated detoxified sentences are then checked for the absence of toxicity and semantic similarity to the original sentence.

We use Russian toxic sentences from the corpora of user utterances taken from Russian social networks Odnoklassniki (Kaggle, 2019) and Pikabu (Kaggle, 2020), and from the Russian segment of Twitter (Rubtsova, 2015). We select only the sentences which were classified as toxic by a pretrained toxicity classifier. The classifier is a ruBERT model (Kuratov and Arkhipov, 2019) fine-tuned on the same datasets. Overall, our dataset contains 8,622 sentences. We use 6,947 of them as training data, 800 for validation and 875 for testing models.

3.2 Competition Rules

The competition rules allowed the participants to use the collected dataset and any additional corpora and pre-trained models as long as they are free and publicly available. The participants could also use our baseline models in any way.

We evaluated the participating models both manually and automatically on the test set. We used state-of-the-art techniques for both evaluations. Due to the large amount of manual evaluation we resort to crowdsourcing instead of expert annotation.

4 Detoxification Models

4.1 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 applies it to input sentences.

RuT5 Baseline Another approach is the supervised baseline based on the T5 model. We fine-tune the ruT5-base model[5] on the train part of the provided dataset.

[3] https://russe.nlpub.org/2022/tox

[4] https://toloka.yandex.ru/en

[5] https://huggingface.co/sberbank-ai/ruT5-base

RuPrompts The third baseline is based on the library ruPrompts[6] for fast language model tuning via automatic prompt search. The method Continuous Prompt Tuning (Konodyuk and Tikhonova, 2021) is to train with gradient descent embeddings corresponding to the prompts, such approach is less expensive to compare with classic fine-tuning of a big language models. In the baseline we tuned the prompts for the ruGPT3-large model. Pre-trained prompts for the baseline is available in huggingface[7].

4.2 Participants

We briefly describe the models developed by participants. More details about the participating systems can be found in (Dementieva et al., 2022)

Team 1 (ruT5-finetune) Authors approach is based on the ruT5 model[8]. 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.

Team 2 (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.

Team 3 (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[9] for the sequence tagging task. The ruT5-base model was trained to fill in the tokens for coarse-grain edit type *replace*.

Team 4 (ruGPT3-XL) trained RuGPT3 XL[10] to generate a non-toxic text on the competition train data. The input is the concatenation of the toxic and non-toxic sentences.

Team 5 (RoBERTa-replace) solution is based on the RoBERTa-large[11]. 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.

Team 6 (ruT5-clean) used the ruT5-large model[12] improved by data cleaning. The preprocessing stage consitsts of emoticons and smiley filtering and removing duplicate characters. The Levenshtein Transformer (Susanto et al., 2020) was used as an extra step in preprocessing to clean the ruT5-large model output.

Team 7 (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.

Team 8 (ruT5-preproc) This solution is based on ruT5-base model with additional pre- and post-processing of the texts. Team finetuned the ruT5-base model on the provided data and used heuristics for text pre/processing.

Team 9 (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.

Team 10 (ruPrompts-plus) This team advanced over the ruPrompts baseline. The solution is based on RuGPT3-XL (Generative Pretrained Transformer-3 for Russian) [13] adapted to the task via prompt tuning. Using RuGPT3-XL as a frozen backbone, team trains only a sequence of continuous embeddings inserted before and after an input text.

[6] https://sberbank-ai.github.io/ru-prompts

[7] https://huggingface.co/konodyuk/prompt_rugpt3large _detox_russe

[8] https://huggingface.co/sberbank-ai/ruT5-base

[9] https://huggingface.co/DeepPavlov/rubert-base-cased-conversational

[10] https://huggingface.co/sberbank-ai/rugpt3xl

[11] https://huggingface.co/sberbank-ai/ruRoberta-large

[12] https://huggingface.co/sberbank-ai/ruT5-large

[13] https://huggingface.co/sberbank-ai/rugpt3xl

5 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 a text, content preservation, and fluency of a text. The three metrics are then aggregated to a joint score. We use the following techniques.

Style ($\mathbf{STA}_a$) is evaluated with a BERT-based classifier for toxicity detection. We use the same ruBERT-based classifier that was used for pre-selection (see Section 3.1).

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, as 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.

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., 2018). 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 it easier to combine the three metrics easier.

Joint ($\mathbf{J}_a$) Following Krishna et al. (2020), we combine the three metrics at the sentence level by multiplying them. The document-level score is computed as the average of scores for all sentences.

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.

6 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.

6.1 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 ($\mathbf{STA}_m$) 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 meaningless 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 ($\mathbf{J}_m$) We aggregate the three metrics by multiplying sentence-level scores. Since all scores are binary, the joint score is 1 only if all three metrics are 1. Therefore, it indicates fully acceptable sentences.

6.2 Crowdsourcing Setup

Each of the three parameters is evaluated in a separate crowdsourcing project. For all the projects, the evaluation was made by only native Russian speakers.

6.2.1 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 (see Figure 2). Finally, we apply the fluency evaluation task (see Figure 3) to both the source and the target and compute the final fluency score from the source and target scores.

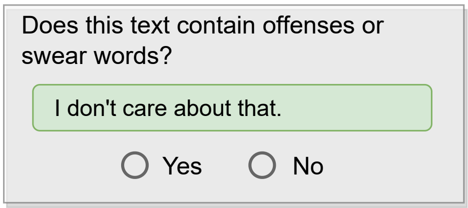
Each sentence in each of the projects is labelled by 10 to 12 workers. We aggregate their result using Dawid-Skene aggregation method (Dawid and Skene, 1979). It takes into account the dynamically defined reliability of workers. For each example with multiple labels Dawid-Skene method returns the label 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.

6.2.2 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

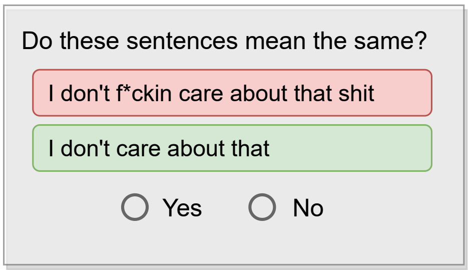


Does this text contain offenses or swear words?

I don't care about that.

○ Yes ○ No

Figure 1: Interface of the toxicity detection task.



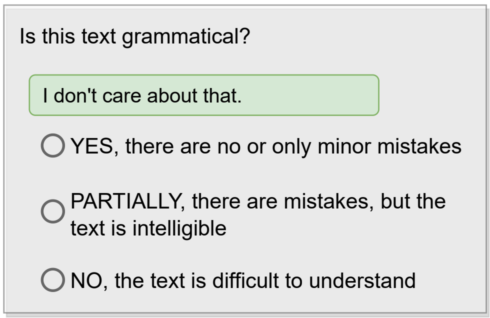
Do these sentences mean the same?

I don't f*ckin care about that shit

I don't care about that

○ Yes ○ No

Figure 2: Interface of the content similarity task.



Is this text grammatical?

I don't care about that.

○ YES, there are no or only minor mistakes

○ PARTIALLY, there are mistakes, but the text is intelligible

○ NO, the text is difficult to understand

Figure 3: Interface of the fluency evaluation task.

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).

	STA$_a$	SIM$_a$	FL$_a$	J$_a$	ChrF
adversarial	0.97	**0.94**	0.96	**0.87**	0.53
ruT5-finetune	**0.98**	0.86	**0.97**	0.82	0.55
ruT5-large	0.95	0.86	**0.97**	0.78	0.57
ruT5-clean	0.95	0.82	0.91	0.71	0.57
lewis	0.93	0.80	0.88	0.66	0.56
ruGPT3-XL	0.94	0.73	0.89	0.61	0.50
RuT5 Baseline	0.80	0.83	0.84	0.56	0.57
ruPrompts-plus	0.80	0.80	0.83	0.54	0.56
ruPrompts	0.81	0.79	0.80	0.53	0.55
ruT5-preproc	0.85	0.76	0.78	0.52	0.53
human references	0.85	0.72	0.78	0.49	**0.77**
ruGPT3-filter	0.83	0.76	0.76	0.48	0.51
RoBERTa-replace	0.57	0.89	0.91	0.44	0.54
Delete	0.56	0.89	0.85	0.41	0.53
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.

7 Results

In this section, first we present the data, namely the outcome of the shared task on detoxification evaluation. Second, we perform anlysis of correspondance of human and automatic metics. Finally, we conclude with a discussion of assessors's performance and overall difficulty of the task.

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 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. 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.

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 one of ruT5-based systems.

	STA$_m$	SIM$_m$	FL$_m$	J$_m$
human references	**0.89**	0.82	**0.89**	**0.65**
ruT5-clean	0.79	**0.87**	**0.90**	0.63
RuT5 Baseline	0.79	0.82	**0.92**	0.61
ruT5-large	0.73	**0.87**	**0.92**	0.60
lewis	0.82	0.79	0.85	0.58
ruPrompts-plus	0.78	0.81	**0.90**	0.57
ruT5-finetune	0.80	0.78	0.87	0.56
ruT5-preproc	0.79	0.72	0.78	0.51
ruGPT3-XL	0.81	0.70	0.90	0.50
ruPrompts	0.80	0.70	**0.87**	0.49
ruGPT3-filter	0.77	0.72	0.83	0.45
RoBERTa-replace	0.43	0.62	0.79	0.17
Delete	0.39	0.71	0.73	0.16
Duplicate	0.11	1.00	1.00	0.11
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 figures **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).

Metric	STA$_a$	SIM$_a$	FL$_a$	J$_a$	ChrF
STA$_m$	**0.695**	**-0.888**	-0.398	0.305	0.264
SIM$_m$	-0.305	-0.153	-0.042	-0.431	0.276
FL$_m$	-0.237	-0.291	-0.116	-0.425	0.218
J$_m$	**0.595**	**-0.746**	-0.380	0.278	0.367

Table 4: Pearson's correlation coefficient between automatic VS manual metrics on system level. Bold numbers denote the statistically significant correlation (p-value ≤ 0.05).

However, ruT5-clean (the best-performing participant) is not significantly better than the ruT5 baseline. Interestingly, the **adversarial** model whose automatic scores are the highest, in fact produces sentences of an very low quality.

7.2 Automatic vs Manual Metrics

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

We check this assumption by computing the Spearman ρ correlations at three different levels: sentence level, system level and system ranking

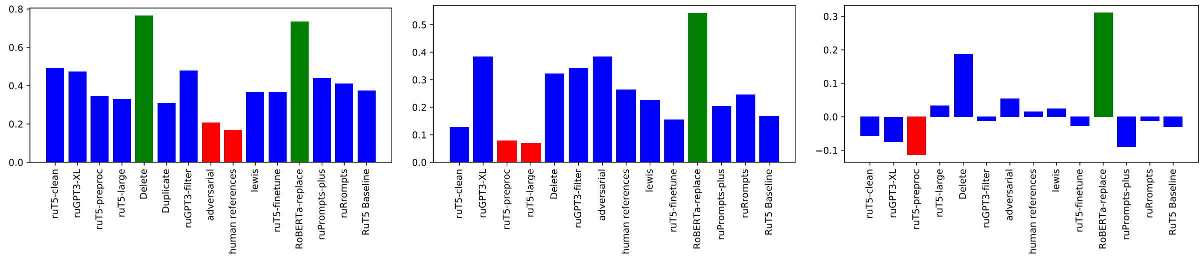


Figure 4: Correlations between automatic and manual metrics at the sentence level for different models. (Right: **STA** metric; Center: **SIM** metric; Left: **FL** metric.)

Metric	STA$_a$	SIM$_a$	FL$_a$	J$_a$
STA$_m$	-0.437	**0.679**	0.226	0.345
SIM$_m$	0.187	-0.126	0.099	0.022
FL$_m$	0.165	-0.314	0.037	-0.046
J$_m$	-0.041	0.020	0.275	0.178

Table 5: Spearman's correlation coefficient between automatic VS manual metrics based on system ranking. Bold numbers denote the statistically significant correlation (p-value ≤ 0.05).

Metric	BertScore	ROUGE-L	BLEU	ChrF
STA$_m$	**-0.710**	**-0.550**	**-0.600**	-0.296
SIM$_m$	**0.819**	**0.802**	**0.863**	0.495
Fl$_m$	**0.796**	**0.675**	**0.700**	0.464
J$_m$	**0.661**	**0.657**	**0.546**	0.325

Table 6: Spearman's correlation coefficient between automatic style transfer VS manual metrics based on system ranking. Bold numbers denote the statistically significant correlation (p-value ≤ 0.05).

level. At sentence level, we compare automatic metrics for each sentence and then compare them across their manual analogies. For the system level we first compute average scores for each participant and each metric and them uses such vectors of scores to calculate correlations. As for the system ranking level, we use the rank of the system in the ranked system list instead of the scores, which allows to not take the difference of score distributions into account. The last metric is trying to assess the capability of a metric to predict the outcome of a competition.

7.2.1 System Level Correlations

At the system level we compute correlation scores of all metrics. We highlight all high correlations (the absolute value above 0.6) in Table 4. We clearly see that none of automatic metrics correlate with their manually measured counterparts. On the other hand, there is strong negative correlation between the manual style and automatic content preservation score.

Moreover, manual content and fluency 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., 2021a).

7.2.2 System Ranking Level Correlations

We also compute the correlation of rankings of models produced by different metrics using Spearman's ρ correlation. According to Table 5, we mostly see weak or no correlation. The rankings by automatic metrics of style, content preservation, and fluency do not correlate with their counterparts produced by manual metrics, apart from the correlation of manual metric of style evaluation (STA$_m$) and automatic metric of content preservation (SIM$_a$).

Despite that ChrF metric counted as more suitable text generation metric for the Russian Language, additionally we computed correlations for other text generation metrics as BLEU (Papineni et al., 2002), ROUGE-L (Sutherlin et al., 2011), and BertScore (Zhang et al., 2020a). The results are presented in the Table 6. Unexpectedly, ChrF does not correlated at all with the manually computed manual metrics, according to the ranking evaluation. BertScore, ROUGE-L, BLEU demonstrated quite strong correlations with the manual metrics, which are statistically significant in comparison to the ChrF scores. At the same time, from the Table 6 we can conclude that even the highest correlation numbers (0.661) in our case cannot guarantee high-quality prediction of manual metrics, which still requires further manual evaluation steps.

7.2.3 Sentence-level Correlations

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 4). We see that automatic and manual toxicity scores are much better correlated for the **Delete** and **RoBERTa-replace** 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, **adversarial** 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 **RoBERTa-replace** model which leaves the majority of words intact, so for it similarity boils down to word matching. Instead, 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 (**RoBERTa-replace** and **Delete**). Overall, we see that it is more difficult to correctly classify *better-performing models* and *models based on large pre-trained language models*. This suggests that the automatic evaluation might fail exactly where we need it most, i.e. in discriminating between the good models.

7.3 Assessors Performance

While in many works the human evaluation is considered as undoubtedly reliable, we notice that this is not always true. Human evaluation can suffer from: (i) the low reliability of crowd workers and (ii) the difficulty and subjectivity of the tasks.

In crowdsourcing experiments, it is common to give each example for labelling to 3–5 people and aggregate the labels. It our case 3 annotations per sample were not enough. They yielded a labelling with around 10% mistakes. Thus, we collected 10 annotations per sample. Such labelling was more reliable: the error rate did not exceed 3% for style and content and 6% for fluency.

To measure the difficulty of the task, we compute inter-annotator agreement coefficient Krippendorff's alpha (Krippendorff, 2011). It turns out that the agreement is moderate: content: 0.522, 0.448, and 0.394 for style, content, and fluency, respectively. The expert Krippendorff's alpha scores are close: 0.584, 0.458, and 0.463. This confirms that in the experiment with 10 annotations per example the crowd workers are reliable enough, but the task itself is subjective.

Interestingly, the style evaluation gains the highest inter-annotator agreement, just as it had the highest correlation between the manual and the automatic labelling. This suggests that that toxicity is more stable and better interpreted by both humans and models.

8 Conclusion

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. At the system ranking level BertScore metric yielded the best correlation with human judgements.

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 detoxification task (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. A worse thing is that the automatic metrics produce less reliable for better-performing models, thus blocking the advance of style transfer models.

Acknowledgements

This work was supported by MTS-Skoltech laboratory on AI.

References

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

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021b. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online. Association for Computational Linguistics.

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2915–2924, Online. Association for Computational Linguistics.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Daryna Dementieva, Irina Nikishina, Varvara Logacheva, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. RUSSE-2022: Findings of the First Russian Detoxification Task Based on Parallel Corpora. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*.

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

Fei Huang, Zikai Chen, Chen Henry Wu, Qihan Guo, Xiaoyan Zhu, and Minlie Huang. 2021. NAST: A non-autoregressive generator with word alignment for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1577–1590, Online. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Kaggle. 2019. Russian language toxic comments. `https://www.kaggle.com/blackmoon/russian-language-toxic-comments`. Accessed: 2021-03-01.

Kaggle. 2020. Toxic russian comments. `https://www.kaggle.com/alexandersemiletov/toxic-russian-comments`. Accessed: 2021-03-01.

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

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

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

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

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021a. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021b. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102, Online. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Kevin Lin, Ming-Yu Liu, Ming-Ting Sun, and Jan Kautz. 2020. Learning to generate multiple style transfer outputs for an input sentence. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 10–23, Online. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with Parallel Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.

Yun Ma, Yangbin Chen, Xudong Mao, and Qing Li. 2021. Collaborative learning of bidirectional decoders for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9250–9266, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 138–147, Hong Kong. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3786–3800, Online. Association for Computational Linguistics.

Yu. Rubtsova. 2015. Rutweetcorp. `https://study.mokoron.com/`. Accessed: 2022-03-01.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. “transforming” delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

Daniel P Sutherlin, Linda Bao, Megan Berry, Georgette Castanedo, Irina Chuckowree, Jenna Dotson, Adrian Folks, Lori Friedman, Richard Goldsmith, Janet Gunzner, et al. 2011. Discovery of a potent, selective, and orally available class i phosphatidylinositol 3-kinase (pi3k)/mammalian target of rapamycin (mtor) kinase inhibitor (gdc-0980) for the treatment of cancer. *Journal of medicinal chemistry*, 54(21):7579–7587.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yi Zhang, Tao Ge, and Xu Sun. 2020b. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7135–7144, Online. Association for Computational Linguistics.

Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2021. Multimodal text style transfer for outdoor vision-and-language navigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1207–1221, Online. Association for Computational Linguistics.