Text Detoxification using Large Pre-trained Neural Models

David Dale[‡], **Anton Voronov**[‡,†], **Daryna Dementieva**[‡], **Varvara Logacheva**[‡], **Olga Kozlova**[†], **Nikita Semenov**[†], **and Alexander Panchenko**[‡]
[‡]Skolkovo Institute of Science and Technology, Moscow, Russia
[†]Mobile TeleSystems (MTS), Moscow, Russia
{d.dale,anton.voronov,daryna.dementieva,v.logacheva,a.panchenko}@skoltech.ru
{oskozlo9,nikita.semenov}@mts.ru

Abstract

We present two novel unsupervised methods for eliminating toxicity in text. Our first method combines two recent ideas: (1) guidance of the generation process with small style-conditional language models and (2) use of paraphrasing models to perform style transfer. We use a well-performing paraphraser guided by style-trained language models to keep the text content and remove toxicity. Our second method uses BERT to replace toxic words with their non-offensive synonyms. We make the method more flexible by enabling BERT to replace mask tokens with a variable number of words. Finally, we present the first large-scale comparative study of style transfer models on the task of toxicity removal. We compare our models with a number of methods for style transfer. The models are evaluated in a reference-free way using a combination of unsupervised style transfer metrics. Both methods we suggest yield new SOTA results.

1 Introduction

Identification of toxicity in user texts is an active area of research (Zampieri et al., 2020; D'Sa et al., 2020; Han and Tsvetkov, 2020). The task of automatic rewriting of offensive content attracted less attention, yet it may find various useful applications such as making online world a better place by suggesting to a user posting a more neutral version of an emotional comment. The existing works on text detoxification (dos Santos et al., 2018; Tran et al., 2020; Laugier et al., 2021) cast this task as style transfer. The style transfer task is generally understood as rewriting of text with the same content and with altering of one or several attributes which constitute the "style", such as authorship (Voigt et al., 2018), sentiment (Shen et al., 2017), or degree of politeness (Madaan et al., 2020). Despite the goal of preserving the content, in many cases changing the style attributes changes the meaning of a sentence significantly.[1] So in fact the goal of many style transfer models is to transform a sentence into a somewhat similar sentence of a different style on the same topic.[2] We suggest that detoxification needs better preservation of the original meaning than many other style transfer tasks, such as sentiment transfer, so it should be performed differently.

We present two models for text detoxification, which have extra control for content preservation. The first model, **ParaGeDi**, is capable of fully regenerating the input. It is based on two ideas: external control of an output of a generation model by a class-conditioned LM (Krause et al., 2020) and formulation of style transfer task as paraphrasing (Krishna et al., 2020). Being based on a paraphraser model, **ParaGeDi** explicitly aims at preserving the meaning of the original sentence. The second approach, **CondBERT**, inspired by Wu et al. (2019a), follows the pointwise editing setup. It uses BERT to replace toxic spans found in the sentence with their non-toxic alternatives. The semantic similarity is maintained by showing the original text to BERT and reranking its hypotheses based on the similarity between the original words and their substitutes. Interestingly, BERT does not need any class-conditional pre-training to successfully change the text style from toxic to normal.

In addition, we perform a large-scale evaluation of style transfer models on detoxification task, comparing our new models with baselines and state-of-the-art approaches. We release our code and data.[3]

Our contributions are as follows:

- We propose two novel detoxification methods based on pre-trained neural language models: **ParaGeDi** (paraphrasing GeDi) and **CondBERT** (conditional BERT).

[1] For example, Lample et al. (2019) provide the following sentence as an example of transfer from male to female writing: *Gotta say that beard makes you look like a Viking → Gotta say that hair makes you look like a Mermaid.*

[2] A formal task definition is presented in Appendix A.

[3] https://github.com/skoltech-nlp/detox

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7979–7996
November 7–11, 2021.

- We conduct an evaluation of these models and their comparison with a number of state-of-the-art models for text detoxification and sentiment transfer and release the detoxification dataset.
- We create an English parallel corpus for the detoxification task by retrieving toxic/safe sentence pairs from the ParaNMT dataset (Wieting and Gimpel, 2018). We show that it can further improve our best-performing models.

2 Related Work

One of the most straightforward ways of solving style transfer task is to "translate" a source sentence into the target style using a supervised encoder-decoder model (Rao and Tetreault, 2018). Since the source and the target are in the same language, pre-trained LMs such as GPT-2 (Radford et al., 2019) can be applied for this task — fine-tuning them on relatively small parallel corpora gives a good result (Wang et al., 2019). However, this method is used quite rarely because of the lack of sufficiently large parallel data. The rest of described models are trained in an unsupervised way.

Pointwise Editing Models A relatively easy yet efficient style transfer method is to leave the sentence intact and manipulate only individual words associated with the style. Delete-Retrieve-Generate (DRG) framework (Li et al., 2018) was the first effort to perform such transfer. It proposes four methods based on this principle. Two of them perform well on our data: **DRG-RetrieveOnly** retrieves a sentence with the opposite style which is similar to the original sentence and returns it, and **DRG-TemplateBased** takes the style attributes from it and plugs them into the original sentence. Here, the performance depends on the methods for the identification of style markers and retrieval of replacements. Words associated with style are typically identified either based on their frequencies as in the original paper, some works use attention weights as features (Sudhakar et al., 2019).

Alternatively, style transfer can use Masked Language Modelling (MLM). An MLM trained on a dataset with style labels picks a replacement word based not only on the context, but also on the style label. An example of such model is **Mask & Infill** (Wu et al., 2019b). It is most similar to **CondBERT** method we propose. However, CondBERT performs additional control over the style and the content preservation and is able to make multi-word replacements. Another similar model of this type is described by Malmi et al. (2020). It has a more complicated structure: there, two MLMs trained on corpora of different styles perform replacements jointly.

End-to-end Architectures In contrast to these models, there exist end-to-end architectures for style transfer. They encode the source sentence, then manipulate the resulting hidden representation in order to incorporate a new style, and then decode it. Some of them disentangle the hidden representation into the representation of content and style (John et al., 2019). The others force the encoder to represent style-independent content (Hu et al., 2017). Alternatively, the model **DualRL** by Luo et al. (2019) performs a direct transfer from the source to the target style. The task is paired with the dual task (back transfer to the source style) which allows models to train without parallel data. The Deep Latent Sequence Model (**DLSM**) model by He et al. (2020) uses amortized variational inference to jointly train models for the primal and dual tasks. The Stable Style Transformer (**SST**) method (Lee, 2020) trains a pair of sequence-to-sequence transformers for primal and dual tasks using cross-entropy of a pretrained style classifier as an additional discriminative loss. The Style Transfer as Paraphrase (**STRAP**) method by Krishna et al. (2020) views a style transfer model as a paraphraser that adds attributes of a particular style to a text. The authors create pseudo-parallel data by transferring style-marked texts to neutral with a pre-trained general-purpose paraphraser and then train sequence-to-sequence models on these neutral-to-styled parallel datasets. Our **ParaGeDi** model is conceptually similar to these methods. However, unlike these methods, the style is not infused into the model or a sentence representation but is imposed on the generator by another model.

Detoxification Detoxification of text is a relatively new style transfer task. The first work on this topic by (dos Santos et al., 2018) is an end-to-end seq2seq model trained on a non-parallel corpus with autoencoder loss, style classification loss and cycle-consistency loss. A more recent work by Tran et al. (2020) uses a pipeline of models: a search engine finds non-toxic sentences similar to the given toxic ones, an MLM fills the gaps that were not matched in the found sentences, and a seq2seq model edits the generated sentence to make it more fluent. Finally, Laugier et al. (2021)

detoxify sentences by fine-tuning T5 as a denoising autoencoder with additional cycle-consistency loss. Dathathri et al. (2020) and Krause et al. (2020) approach a similar problem: preventing a language model from generating toxic text. They do not need to preserve the meaning of the input text. However, the idea of applying a discriminator to control an LM during generation can be used for style transfer, as we show in our experiments.

3 Paraphrasing GeDi Model

The recently proposed GeDi model (Krause et al., 2020) performs text generation from scratch guided by a language model informed about specific attributes of a text, e.g. style or topic. We extend this model by enabling it to paraphrase the input text.

3.1 GeDi

The original GeDi model consists of two components: a generation model (GPT-2) and a discrimination model, which is also a GPT-2 trained on sentences with additional sentence-level style labeling — during training the style label is prepended to a sentence. This makes the discriminating model learn the word distributions conditioned on a particular label. At each generation step, the distribution of the next token predicted by the main model P_{LM} is modified using an additional class-conditional language model P_D and the Bayes rule:

$$P(x_t|x_{<t}, c) \propto P_{LM}(x_t|x_{<t})P_D(c|x_t, x_{<t})$$

Here, x_t is the current token, $x_{<t}$ is the prefix of the text, and c is the desired attribute (e.g. toxicity or sentiment) — one of C classes. The first term is produced by the main language model P_{LM}, and the second term is calculated using the Bayes rule and the additional class-conditional language model P_{CC}. Thus, the tokens which are more likely to appear in a text of the chosen style get a higher probability:

$$P_D(c|x_t, x_{<t}) \propto P(c)P_{CC}(x, x_{<t}|c)$$

The name GeDi stands for Generative Discriminator, because a language model, which is generative by its nature, is used as a discriminator for guiding the generation process. GeDi was successfully applied to guiding a GPT-2 language model towards generating texts of particular topics and making the generated text less toxic.

3.2 ParaGeDi

In order to enable GeDi to preserve the meaning of the input text, we replace the regular language model in it with a model capable of paraphrasing. If we denote the original text by x, the generated text of length T by y, and the desired style by c, ParaGeDi models the following probability:

$$\begin{aligned} P(y_t|y_{<t}, x, c) &\propto P_{LM}(y_t|y_{<t}, x)P(c|y_t, y_{<t}, x) \\ &\approx P_{LM}(y_t|y_{<t}, x)P_D(c|y_t, y_{<t}) \end{aligned}$$

The last step is an approximation because the class probability should be conditioned on both x and y. However, this approximation, although not being fully justified, allows us to decouple the paraphraser model (which requires a parallel corpus for training) from the style model (which requires only texts with style labels, not necessarily parallel). The paraphraser and the style model can be trained independently. Moreover, we can plug in any paraphraser as long as it shares the vocabulary with the class-conditional LM. The third (optional) component of this model is a reranker — an external model which reweighs the hypotheses generated by the style LM-guided paraphraser with respect to the style. Our reranker is a pre-trained toxicity classifier which chooses the least toxic hypothesis generated by the ParaGeDi model. Figure 1 illustrates the workflow of our model.

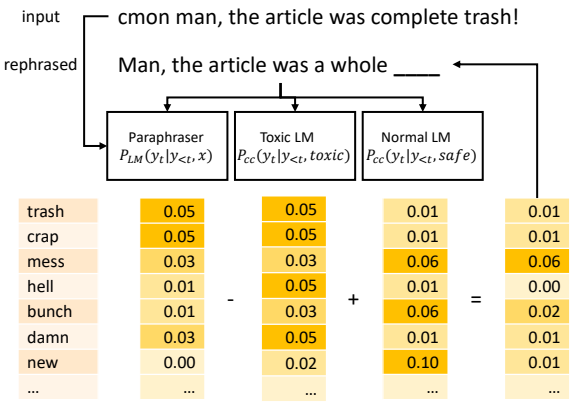

Figure 1: The overview of ParaGeDi model.

ParaGeDi is trained as follows. Its loss $\mathcal{L}_{ParaGeDi}$ consists of a linear combination of two losses: the generative loss $\mathcal{L}_G$ used in LM training, and the discriminative loss $\mathcal{L}_D$ which further pushes different classes away from one another.

$$\mathcal{L}_G = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log P(y_t^{(i)}|y_{<t}^{(i)}, c^{(i)})$$

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N}\log P(c^{(i)}|y_{1:T_i}^{(i)})$$

$$\mathcal{L}_{ParaGeDi} = \lambda\mathcal{L}_D + (1-\lambda)\mathcal{L}_G$$

We enhance the model with a number of inference heuristics that improve content preservation and increase the style transfer accuracy. First, we use a heuristic from the original GeDi model. We raise the conditional LM probability to the power $w > 1$, which biases the discriminator towards the correct class during generation:

$$P(y_t|y_{<t}, x, c) \propto P_{LM}(y_t|y_{<t}, x)P_{CC}(c|y_t, y_{<t})^w$$

Besides that, we suggest two new heuristics:

Smoothing of probabilities — adding a small $\alpha > 0$ to all probabilities from the conditional language model discourages the generation of tokens with low probability conditional on all classes:

$$P_\alpha(c|x_t, x_{<t}) = \frac{\alpha + P(c)P_{CC}(x, x_{<t}|c)}{\sum_{c' \in C}(\alpha + P(c')P_{CC}(x, x_{<t}|c'))}$$

Asymmetric lower and upper bounds (l and u) for class-conditional corrections:

$$P_{\alpha,l,u}(c|x_t, x_{<t}) = \max(l, \min(u, P_\alpha(c|x_t, x_{<t}))).$$

By decreasing the value of u we discourage the insertion of new tokens, as opposed to prohibiting existing tokens. For the problem of detoxification, it means that the model will try less to insert polite words than to delete toxic words from the sentence.

4 Conditional BERT Model

BERT (Devlin et al., 2019) has been trained on the task of filling in gaps ("masked LM"), we can use it to insert non-toxic words instead of the toxic ones. This approach has been suggested by Wu et al. (2019a) as a method of data augmentation. The authors identify words belonging to the source style, replace them with the `[MASK]` token, and the BERT model then inserts new words of the desired style in the designated places. To push BERT towards the needed style, the authors fine-tune BERT on a style-labelled dataset by replacing segmentation embeddings of original BERT with trainable style embeddings.

We perform some changes to this model to adapt it for the detoxification task. While in the original conditional BERT model the words are masked randomly, we select the words associated with toxicity. This can be done in different ways, e.g. by training a word-level toxicity classifier or manually creating a vocabulary of rude and toxic words. We use a method which does not require any additional data or human effort. We train a logistic bag-of-words toxicity classifier. This is a logistic regression model which classifies sentences as toxic or neutral and uses their words as features. As a byproduct of the training process, each feature (word) yields a weight which roughly corresponds to its importance for classification. The words with the highest weights are usually toxic. We use the normalised weights from the classifier as toxicity score. The overview of CondBERT is shown in Figure 2.

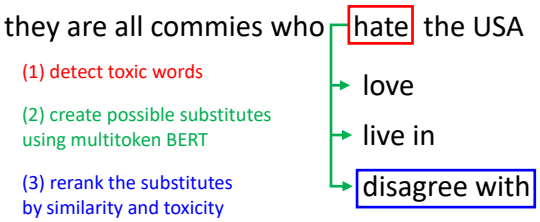

Figure 2: The overview of the CondBERT model.

For each word in a sentence, we compute the toxicity score and then define toxic words as the words with the score above a threshold $t = max(t_{min}, max(s_1, s_2, ..., s_n)/2)$, where $s_1, s_2, ..., s_n$ are scores of all words in a sentence and $t_{min} = 0.2$ is a minimum toxicity score. This adaptive threshold allows balancing the percentage of toxic words in a sentence so that we avoid cases when too many or no words are marked as toxic.

To preserve the meaning of the replaced word, we employ the content preservation heuristics suggested by Arefyev et al. (2020): (i) Preserve the original tokens instead of masking them before the replacement; (ii) Rerank the replacement words suggested by BERT by the similarity of their embedding with the embedding of the original word.

Despite using class-specific sentence embeddings, conditional BERT often predicts toxic words, apparently paying more attention to the context than to the embeddings of the desired class. To force the model to generate non-toxic words we calculate the toxicity of each token in BERT vocabulary and penalize the predicted probabilities of tokens with positive toxicities.

Finally, we enable BERT to replace a single `[MASK]` token with multiple tokens. We generate each next token progressively by beam search and score each multitoken sequence by the harmonic mean of the probabilities of its tokens.

5 Detoxification Experiments

We train the two new models and a number of other systems for text detoxification. Below we describe datasets, evaluation setup, and results.

5.1 Toxicity Classifier

We train two binary classifiers of toxicity. One of them is used to rerank hypotheses in the ParaGeDi model, and the other participates in the evaluation. We train these two classifiers on different sets of data. The overall dataset is the merge of the English parts of the three datasets by Jigsaw (Jigsaw, 2018, 2019, 2020), containing around 2 million examples. We split it into two parts and fine-tune a RoBERTa model (Liu et al., 2019) on it. We use the `roberta-large` model from the original repository. The classifiers perform closely on the test set of the first Jigsaw competition, reaching the AUC-ROC of 0.98 and F_1-score of 0.76.

5.2 Dataset

For training and testing of the style transfer models, we use the English data from the first Jigsaw competition (Jigsaw, 2018). The majority of our methods are trained on non-parallel corpora of source and target styles. To prepare the toxic dataset, we divide the comments labelled as toxic into sentences (the original comments are often too long) and classify each of them with our toxicity classifier. Sentences classified as toxic are used as the toxic part of the dataset (we find 154,771 of them). To select the neutral part of the dataset, we randomly pick the same number of non-toxic sentences from the sentence-separated Jigsaw data. The test set is prepared analogously to the test set of the Jigsaw competition: we use 10,000 sentences with the highest toxicity score according to our classifier.

5.3 Metrics

There is no parallel test set available for the detoxification task, so we cannot use BLEU, METEOR or ROUGE metrics and resort to referenceless evaluation. Style transfer models need to change the style, preserve content and produce a fluent text. These parameters are often inversely correlated, so we need a compound metric to find a balance between them. We follow the evaluation strategy of Krishna et al. (2020) and use the metric **J**, which is the multiplication of sentence-level *style accuracy*, *content preservation*, and *fluency*. The system-level **J** is the average of sentence-level scores. *Style accuracy* (ACC) is measured with a pre-trained toxicity classifier described in Section 5.1. *Content preservation* (SIM) is evaluated as the similarity of sentence-level embeddings of the original and transformed texts computed by the model of Wieting et al. (2019). *Fluency* (FL) measured with the classifier of linguistic acceptability trained on the CoLA dataset (Warstadt et al., 2019). **J** is computed as the average of their sentence-level product. In addition to that, we tried a similar aggregated metric **GM** (Pang and Gimpel, 2019; Laugier et al., 2021) which uses perplexity as the measure of fluency and employs a different aggregation method. Our preliminary experiments showed that **J** and **GM** are strongly correlated, so we keep only **J** for further evaluation.

5.4 Implementation Details

For ParaGeDi, we use a pre-trained T5-based (Raffel et al., 2020) paraphraser,[4] fine-tuned on a random subsample of the ParaNMT dataset (Wieting and Gimpel, 2018). As a discriminator, we fine-tune the `gpt2-medium` model (Radford et al., 2019) on the training part of the Jigsaw-1 dataset using two control codes for toxic and polite texts. Before fine-tuning, we change the vocabulary of the discriminator to match that of T5, and update its embeddings accordingly. We train the discriminator using a combined generative and discriminative loss from Krause et al. (2020), adapting their code for this purpose.

We use beam search decoding with 10 beams to generate paraphrase candidates with the paraphraser and discriminator described above. We apply the classifier from section 5.1 to select the least toxic candidate from the 10 resulting paraphrases.

5.5 Competing Methods

We compare our models with state-of-the-art methods described in Section 2: DRG-TemplateBased, DRG-RetrieveOnly, Mask&Infill, DLSM, STRAP, and SST. We also implement three other baselines: Machine Translation, Detoxifying GPT-2, and Paraphraser. We do not directly compare our models with GeDi, because it is a language model and was not explicitly trained to transform texts.

Machine Translation There is evidence that automatic translation tends to eliminate toxicity (Prabhumoye et al., 2018). Thus, we use a chain

[4] https://huggingface.co/ceshine/t5-paraphrase-paws-msrp-opinosis

of Machine Translation models for detoxification. Namely, we perform English → Pivot → English translation. We choose French and Igbo as pivot languages. French is resource-rich and structurally similar to English, which ensures high-quality translations. Conversely, Igbo is low-resourced and syntactically different. Both experiments are conducted using Google Translate API.

Detoxifying GPT-2 GPT-2 (Radford et al., 2019) can be adapted to a wide range of NLP tasks using a very small task-specific dataset. We experiment with the model's ability to perform sequence-to-sequence tasks. We train it on a parallel dataset of 200 toxic and safe sentences. We randomly select toxic sentences from the Google Jigsaw toxic comments dataset (Jigsaw, 2018) and manually rewrite them in the neutral tone.

Paraphraser Krishna et al. (2020) suggest that a general-purpose paraphraser can remove style markers from text. We check this assumption.

5.6 Results

The performance of all tested models is given in Table 1. Both **ParaGeDi** and **CondBERT** outperform other models by a large margin. The success of CondBERT is explained by its use of heuristics targeted at the components of the metric: (i) it is penalized for generating toxic tokens, which ensures a high ACC score, (ii) over 80% tokens stay unchanged, and the replacements are selected with respect to the similarity to the original words, increasing the overall SIM score, (iii) MLM is pretrained to replace masked tokens with plausible substitutes, increasing FL. ParaGeDi is behind in terms of similarity but has a slightly higher fluency because generation is a better strategy in terms of text naturalness than pointwise corrections. The closest competitor of our models is Mask&Infill which uses similar principles as CondBERT. However, some engineering decisions (e.g. masking of all words at once) result in a substantial drop in fluency and some decrease in style transfer accuracy.

Surprisingly, many advanced models perform below the simplistic (DRG) models **TemplateBased** and **RetrieveOnly**. TemplateBased achieves a high similarity because it keeps most of the original sentence intact, and RetrieveOnly yields a high similarity and style transfer accuracy, because it outputs real non-toxic sentences from the training data. **DLSM** and **SST** models perform full re-generation of text (as opposed to pointwise corrections). More importantly, their decoders are trained from scratch on a relatively small dataset, hence their low fluency scores. Conversely, **STRAP**, which also generates the sentence, has the access to the larger pseudo-parallel data, resulting in higher fluency.

Another finding is that **MT** has detoxification ability. However, it is inversely correlated with its quality: the En→Ig→En detoxifies 37% of sentences but has low SIM and FL scores. Conversely, En→Fr→En yields a better output which keeps most of the original features, including toxicity. The same applies to the T5 **paraphraser**. On the other hand, the **GPT-2** model can be trained to detoxify even on a very small number of parallel sentences (200 in our experiments). Although it performs below many other models, we suggest that training it on a larger parallel dataset can boost its performance. We show examples of the paraphrases by the best-performing models in Table 2.

Additional examples and qualitative analysis can be found in Appendices F and E, respectively.

5.7 Parameter Selection

Our models use multiple parameters and heuristics. We perform an ablation study to explore their usefulness. It turns out that the crucial features of CondBERT are multiword replacement which ensures high fluency and toxicity penalty which increases style strength. On the other hand, masking of all tokens at once as well as control of similarity do not affect the quality. More details on the CondBERT ablation study are given in Appendix B.

ParaGeDi has only one training hyperparameter λ which controls the strength of its discriminative loss. We discover its value has only a marginal effect on the overall quality: the value of **J** decreases only for $\lambda = 1$ which constitutes the absence of generative loss (see Figure 3). The style strength control influences the style accuracy, whereas the use of word probability upper bound increases the similarity, and the absence of beam search decreases fluency. On the other hand, reranking, beam size, smoothing do not affect the model performance. An ablation study of the ParaGeDi model can be found in Appendix C.

6 Mining a Parallel Detoxifying Corpus

The STRAP model (Krishna et al., 2020) is based on the assumption that a regular paraphraser can transform a stylistically marked text into a neutral text. Although our experiments show that a para-

Model	ACC	SIM	FL	J
CondBERT (ours)	0.94	0.69	0.77	0.50 ± 0.0037*
ParaGeDi (ours)	0.95	0.66	0.80	0.50 ± 0.0032*
Mask&Infill (Wu et al., 2019b)	0.78	0.80	0.49	0.31 ± 0.0041
DRG-TemplateBased (Li et al., 2018)	0.66	0.82	0.59	0.30 ± 0.0041
DRG-RetrieveOnly (Li et al., 2018)	0.93	0.33	0.84	0.26 ± 0.0019
DLSM (He et al., 2020)	0.62	0.72	0.48	0.17 ± 0.0033
Detoxifying GPT-2 (baseline)	0.54	0.48	0.72	0.17 ± 0.0026
STRAP (Krishna et al., 2020)	0.29	0.69	0.80	0.15 ± 0.0027
En→Ig→En MT (baseline)	0.37	0.68	0.57	0.12 ± 0.0025
T5 paraphraser (baseline)	0.15	0.90	0.87	0.11 ± 0.0029
SST (Lee, 2020)	0.80	0.55	0.12	0.05 ± 0.0019
En→Fr→En MT (baseline)	0.06	0.91	0.81	0.04 ± 0.0019

Table 1: Performance of detoxification models. Gray lines denote baselines, green lines denote models suggested in this work. The models are sorted with respect to the aggregated **J** score. The asterisk * denotes the scores that are significantly higher than the third best model (Mask&Infill) with $p < 0.01$, based on the paired *t-test*.

Input	no room for those stupid antics .
ParaGeDi (ours)	There's no space for all those adversities.
CondBERT (ours)	no room for those . . . those bizarre exploits .
Mask&Infill	no room for those compelling antics .
DRG-TemplateBased	no room for those conventions antics .
Input	zoological conspiracy theories (arabisraeli conflict) , by contrast , sucks horribly .
ParaGeDi (ours)	By contrast, the zoological theories of conspiracy (the Arab-Israeli conflict) suffer horribly.
CondBERT (ours)	zoological conspiracy theories (arabisraeli conflict) , by contrast , worked out badly .
Mask&Infill	zoological conspiracy theories (arabisraeli conflict) , by contrast , ends horribly .
DRG-TemplateBased	zoological conspiracy theories (arabisraeli conflict) , by contrast , instrumental horribly .
Input	this whole article is bullshit .
ParaGeDi (ours)	This article's not a good deal.
CondBERT (ours)	this whole article is considered incomplete .
Mask&Infill	this whole article is compelling .
DRG-TemplateBased	this whole article is substantially .

Table 2: Examples of detoxification by different models.

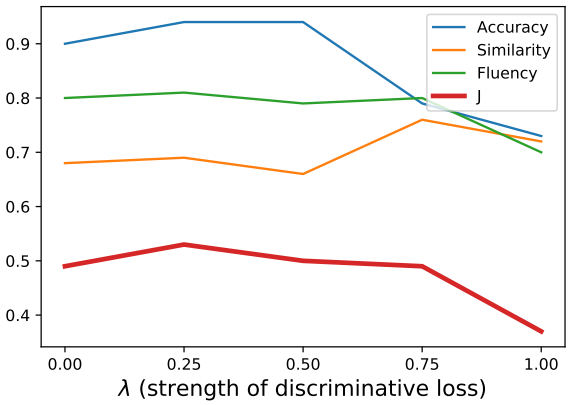

Figure 3: Performance of ParaGeDi with the varying λ parameter (greater λ corresponds to the stronger influence of the discriminative loss and smaller λ means the stronger influence of the generative loss).

phraser, as well as an MT model, are bad detoxifiers on their own (see Section 5.6), we suggest that it is possible to find occasional detoxifying sentence pairs in a large parallel dataset of paraphrases.

Experimental Setup To test this hypothesis, we classify the sentences from the ParaNMT paraphrase dataset (Wieting and Gimpel, 2018) with our toxicity classifier described in Section 5.1 and obtain 500,000 paraphrase pairs where one sentence is more toxic than the other (for more details on the data collection process please see Appendix D). We then compare the regular paraphraser from Section 5.4 fine-tuned on a random subset of ParaNMT (**regular**) with its version fine-tuned on the mined toxic/safe parallel paraphrase corpus (**mined**). We also plug both paraphrasers into ParaGeDi model and compare the overall performance. The results are shown in Table 3.

Model	ACC	SIM	FL	J
Paraphraser				
regular	0.15	0.90	0.87	0.11 ±0.003
mined	0.42	0.87	0.91	0.31 ±0.004
ParaGeDi				
regular	0.94	0.66	0.77	0.50 ±0.003
mined	0.98	0.66	0.84	0.54 ±0.003

Table 3: Comparison of paraphrasers for ParaGeDi.

Discussion of Results None of the paraphrasers

can fully detoxify the test set, but the **mined** paraphraser gets a better ACC than the **regular** one (42% vs 15%). When we replace the regular paraphraser with the detoxifying one in **ParaGeDi**, both detoxification rate and fluency improve without loss in the similarity score. This leaves us with the **J** score of 0.54, which is the highest score we obtained in our detoxification experiment. We do not include it in the main results (Table 1) because this model is not unsupervised. However, this result shows that the general-purpose ParaNMT corpus contains a large number of toxic/safe paraphrase pairs. We believe that mining parallel training sets from large corpora, as opposed to unsupervised methods of style transfer, is a fruitful direction.

7 Human Evaluation of Detoxification

While the automatic reference-free evaluation is cheap and fast, it may be unreliable. Toxicity and fluency classifiers are not perfect and can return erroneous evaluations. The embedding distance which is used to measure the content preservation was shown to weakly correlate with human judgements (Yamshchikov et al., 2021). Thus, we evaluate the best-performing models manually.

Experimental Setup We design our manual evaluation setup to be as close as possible to the automatic evaluation. We evaluate our models along the same three metrics: style accuracy (ACC_m), content similarity (SIM_m), and fluency (FL_m). For all metrics we use a ternary scale: {0, 0.5, 1} corresponding to a bad, partially acceptable, and fully acceptable sentence.

We ask five annotators to evaluate the models. Annotators are NLP researchers with MSc degree or above and with a good command of English. Prior to the annotation, we arranged a preliminary round to reach common annotation understanding. Each sentence is evaluated by three annotators, the final score for a sentence is computed as the average of their scores. We measure the inter-annotator agreement in terms of Krippendorff's α. We obtained the score of 0.42 for the style accuracy, 0.31 for content preservation, and 0.52 for fluency: a moderate agreement for style and fluency annotation, and low agreement for content annotation.

We evaluate three models: our new models ParaGeDi and CondBERT, and Mask&Infill whose automatic scores were the highest among the existing models. The evaluation was conducted on 200 source sentences, each of them was transformed by each of the evaluated models. The input (toxic) sentences for the evaluation were manually preselected to filter out disfluent or senseless utterances (this pre-selection did not consider the outputs). To compensate for the low inter-annotator agreement, we annotate each sample three times and report the average score.

	ACC_m	SIM_m	FL_m	$\mathbf{J_m}$
ParaGeDi (ours)	**93.41**	**64.75**	**91.25**	**55.34**
CondBERT (ours)	91.00	63.92	86.41	50.47
Mask&Infill (top 1)	75.33	59.08	62.08	27.33

Table 4: The results of manual evaluation sorted by $\mathbf{J_m}$. The differences between our models and Mask&Infill are statistically significant with $\alpha < 0.05$ based on the paired *t*-test. Differences between ParaGeDi and CondBERT are significant only for the FL_m metric.

Discussion of Results We show the performance of models in terms of human evaluation in Table 4. The model scores are the averaged sentence scores. We combine the three metrics into a joint quality score which we denote as $\mathbf{J_m}$. Sentence-level $\mathbf{J_m}$ is a multiplication of sentence ACC_m, SIM_m, and FL_m, and the model $\mathbf{J_m}$ scores are the average of sentence scores. This manual evaluation corroborates the superiority of our models over Mask&Infill model. At the same time, it confirms that our two models are not significantly different. Although ParaGeDi outperforms CondBERT in terms of all metrics, the difference in scores is statistically significant only for FL_m.

Besides the evaluation itself, we investigated to what extent the automatic metrics reflect the human judgements. To do that, we compute their Spearman's ρ correlation score with human judgements (see Table 6). For style, we consider the accuracy of toxicity classifier that we used for the evaluation (ACC) and its version which returns the confidence instead of the binary label (ACC-soft). For content we compare SIM (embedding similarity used for computing the **J** score) and BLEU score between the original and detoxified sentence. For fluency, we consider the linguistic acceptability classifier (FL) and perplexity of the GPT-2 (Radford et al., 2019) language model (PPL) which is used for evaluating fluency in many works on style transfer and other generation tasks.

This evaluation shows that the tested metrics of content preservation show only weak correlation with manual scores, which agrees with the previous research (Yamshchikov et al., 2021). The correla-

Model	ACC	SIM	FL	J	BLEU
human	0.81	0.65	0.84	0.445 ± 0.011	1.000
ParaGeDi (ours)	0.93	0.62	0.88	**0.515** ± 0.009*	0.038 ± 0.005
Mask & Infill (Wu et al., 2019b)	0.89	0.76	0.62	0.420 ± 0.013	0.145 ± 0.008
DualRL (Luo et al., 2019)	0.87	0.75	0.63	0.395 ± 0.012	**0.152** ± 0.008
CondBERT (ours)	0.86	0.65	0.62	0.338 ± 0.012	0.125 ± 0.007
SST (Lee, 2020)	0.74	0.65	0.41	0.225 ± 0.011	0.100 ± 0.007
DRG-RetrieveOnly (Li et al., 2018)	0.95	0.29	0.83	0.225 ± 0.006	0.004 ± 0.001
DRG-TemplateBased (Li et al., 2018)	0.82	0.70	0.24	0.115 ± 0.009	0.117 ± 0.007

Table 5: Performance of the sentiment transfer models on the YELP dataset. The models are sorted with respect to the aggregated **J** score. * indicates the score which is significantly higher than the next best model with $p < 0.01$.

$\mathbf{ACC}_m$		$\mathbf{SIM}_m$		$\mathbf{FL}_m$	
ACC-soft	0.59	SIM	0.34	FL	0.54
ACC	0.51	BLEU	0.19	PPL	0.45

Table 6: Spearman's ρ of automatic metrics for evaluating style, content, and fluency with our human scores.

tion of automatic style and fluency metrics with human judgements is moderate. It turns out that the confidence of style classifier is a better style accuracy metric than a binary classifier and the acceptability classifier works better than perplexity, confirming the criticism of perplexity as a fluency metric (Krishna et al., 2020).

8 Sentiment Transfer Experiments

Text detoxification is not as well-established as other style transfer tasks, which makes it is difficult to put our models in the context of other works on style transfer. Thus, we conduct an experiment on a different domain, namely, sentiment transfer.

Experimental Setup We train ParaGeDi and CondBERT on the Yelp reviews dataset (Li et al., 2018) and compare them with Mask&Infill, SST, DRG-TemplateBased, DRG-RetrieveOnly, and DualRL models (see Section 2). We tune the hyperparameters of ParaGeDi and CondBERT on the Yelp development set and use the outputs of other models generated by their authors.

We evaluate the models using the **J** as in our detoxification experiments. For the evaluation of style transfer accuracy, we train two sentiment classifiers on two disjoint parts of the Yelp dataset as in Section 5.1. We use one for inference and another for evaluation. We also compute the BLEU score against human references provided by Li et al. (2018). The results are shown in Table 5, averaged over two transfer directions.

Discussion of Results ParaGedi outperforms other models in terms of **J**. As before, the other models fail to generate fluent texts because they replace only specific words or because they learn to generate texts from scratch. ParaGeDi model is the only competitor which combines pre-trained models and with full regeneration. The performance of the CondBERT model is low on this task, corroborating that detoxification and style transfer for other domains require different techniques.

On the other hand, the BLEU score questions this result. Compared to the human references, the best-performing model is **DualRL** followed by the two MLM-based models: Mask&Infill and our CondBERT. The evaluation of reference human answers also questions the referenceless metrics. First, we see that the ACC score is limited by the classifier performance. Since it gives only 0.81 to presumably 100% correct manually written sentences, the small differences in ACC should not be considered significant, and the ACC above 0.81 is unreliable. Overall, since the score of human answers is close to those of ParaGeDi and Mask&Infill, ParaGeDi can still be considered a strong style transfer model, and more precise evaluation should be done by humans because metrics cannot distinguish between the models at this level.

9 Conclusion

We present two style transfer models tailored for detoxification, i.e. transfer from toxic to non-toxic texts. Both of them combine high-quality pre-trained LMs with the extra style guidance. **ParaGeDi** is based on a paraphraser guided by a style-conditioned GPT-2 model. **CondBERT** model is based on BERT which does not need any fine-tuning, and all style control is performed with a pre-trained toxicity classifier. We conduct a large-scale study of style transfer models exploiting both automatic and manual evaluation. Our experiments show that the proposed methods outperform other state-of-the-art style transfer models on the tasks of detoxification and sentiment transfer.

Ethical Statement

Toxicity is a sensitive topic where the unexpected results and byproducts of research can cause harm. Therefore, we would like to consider some ethical concerns related to our work.

On Definition of Toxicity Toxicity is an umbrella term for almost any undesirable behaviour on the Internet. It ranges from "mild" phenomena like condescending language (Perez Almendros et al., 2020) to grave insults or oppression based on racial or other social-demographic characteristics.

While annotators agree when labelling serious cases of toxicity such as hate speech (Fortuna and Nunes, 2018), the labelling of less severe toxicity is subjective and depends on the annotator's background (Al Kuwatly et al., 2020). This can cause the underestimation of certain types of toxicity. To define the toxicity in the most objective feasible way, we adopt a data-driven approach as presented in detail formally in Appendix A. Both models we propose recognise toxicity based on a toxicity-labelled dataset and do not require any additional manually created dictionaries or rules. Thus, their understanding of toxicity can be tuned with the input data. This ensures that given a corpus with unbiased toxicity labelling our models can produce unbiased detoxification.

On the other hand, in case the training corpus is biased, the model can reproduce the biases, so it should be applied with caution.

Toxification of Texts Detoxification task implies the possibility to perform the opposite transformation, i.e. to rewrite a neutral text into a toxic one. Various style transfer models, including ours, could in principle be used to complete this task. However, in case of CondBERT, the quality of such transformation would be bad, and it would be almost impossible to pass the results of this "toxification" off as real toxic sentences. The reason for that is the structure of toxic data.

One of the main properties of toxic style is the presence of lexical markers of this style (rude or obscene words). Such markers (i) carry most of stylistic information of a sentence (i.e. their presence is a strong indicator of this class), (ii) have synonyms which are free from this stylistic information. Both our methods strongly rely on these properties. They identify toxic words and replace them with non-toxic synonyms. On the other hand, if performing the opposite transformation, we cannot use these properties any more. First, there do not exist non-toxic words which are strong indicators of neutral (non-toxic) style. Second, it is almost infeasible to identify non-toxic words which have toxic synonyms and replace them appropriately. Therefore, we suggest that CondBERT is not suitable for toxification.

The above arguments do not prove that CondBERT or ParaGeDi cannot be applied for toxification. However, they suggest that the quality of the resulting text might not be higher than with simpler toxification methods (e.g. handwritten rules for inserting rude words).

Detoxification as a Censorship Another concern is the fact the detoxification technology could no used to rewrite user-generated messages, which might be considered a form of censorship. We would like to look at that from a different perspective. The social media currently already perform censorship, e.g. Instagram provides tools for removal of messages based on automatically identified harmful content.[5]

On the other hand, we suggest mitigating this policy by rewriting toxic messages instead of removing them altogether. Last but not least, we suggest that user messages should not be modified without user consent. The detoxificaiton models should be used for suggesting detoxifying edits rather than performing them automaticallly.

At the same time, detoxification models can make chatbots safer by detoxifying (if necessary) their answers before sending them to users. An automatically generated toxic comment by a neural chatbot may be the result of pre-training on the biased textual data – a problem which is currently unsolved completely (Gehman et al., 2020). Therefore, a detoxification of automatically generated content might be a valid use-case for minimizing reputational losses for the company created such an unmoderated chatbot (Babakov et al., 2021).

Acknowledgements

This research was conducted under the framework of the joint MTS-Skoltech laboratory. We are grateful to the reviewers for their helpful suggestions which substantially improved this work.

[5] https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nikolay Babakov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Detecting inappropriate messages on sensitive topics that could harm a company's reputation. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 26–36, Kiyv, Ukraine. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 21–25, Marseille, France. European Language Resources Association (ELRA).

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, Online. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Jigsaw. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge. Accessed: 2021-03-01.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification. Accessed: 2021-03-01.

Jigsaw. 2020. Jigsaw multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification. Accessed: 2021-03-01.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 424–434. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *CoRR*, abs/2009.06367.

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.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Leo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. *CoRR*, abs/2102.05456.

Joosung Lee. 2020. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 195–204. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5116–5122. ijcai.org.

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.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality 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 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 189–194. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3267–3277. Association for Computational Linguistics.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards A friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2107–2114. International Committee on Computational Linguistics.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3571–3576. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: training neural machine translation with semantic similarity. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4344–4355. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 451–462. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019a. Conditional BERT contextual augmentation. In *Computational Science - ICCS 2019 - 19th International Conference, Faro, Portugal, June 12-14, 2019, Proceedings, Part IV*, volume 11539 of *Lecture Notes in Computer Science*, pages 84–95. Springer.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5271–5277. ijcai.org.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14213–14220. AAAI Press.

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). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Definition of Text Detoxification Task

In our work, we adhere to the data-driven definition of toxicity. The toxicity is a particular binary value associated with a text: {`toxic`, `neutral`}. We assume that this textual characteristic is measurable using a function $\sigma(x_i) \to s_i$ that obtains as input text x_i and returns the corresponding style label s_i. For instance, it can be implemented using a text classifier.

Let us assume a set of two discreet mutually exclusive styles $S = \{s^{src}, s^{tg}\}$ which corresponds to the source `toxic` and target `neutral` styles. Let us consider two text corpora $D^{src} = \{d_1^{src}, d_2^{src}, ..., d_n^{src}\}$ and $D^{tg} = \{d_1^{tg}, d_2^{tg}, ..., d_m^{tg}\}$ belonging to the source and target styles s^{src} and s^{tg}, respectively. For each text d_i, let us assume that it has a style s_i measurable with the function $\sigma : D \to S$. There also exists a binary function $\delta : D \times D \to [0, 1]$ that indicates the semantic similarity of two input texts and a unary function $\psi : D \to [0, 1]$ that indicates the degree of the text fluency. In general, the sizes of the source and the target corpora D^{src} and D^{tg} are different ($n \neq m$) and the texts in them are not aligned, i.e., in general, $\delta(d_i^{src}, d_i^{tg}) \neq 1$. If $n = m$ and $\delta(d_i^{src}, d_i^{tg}) = 1$ for all texts, this is a special case of a parallel style-aligned corpus. Given the introduced notations, we define the task of text detoxification as follows:

A text detoxification model is a function $\alpha : S \times S \times D \to D$ that, given a source style s^{src}, a target style s^{tg}, and an input text d^{src}, produces an output text d^{tg} such that:

- The style of the text changes from the source style s^{src} to the target style s^{tg}: $\sigma(d^{src}) \neq \sigma(d^{tg}), \sigma(d^{tg}) = s^{tg}$;
- The content of the source text is saved in the target text as much as required for the task: $\delta(d^{src}, d^{tg}) \geq t^{\delta}$;
- The fluency of the target text achieves the required level: $\psi(d^{tg}) \geq t^{\psi}$,

where t^{δ} and t^{ψ} are the error tolerance threshold values for the content preservation (δ) and fluency (ψ) functions.

When removing the toxicity from a text, we inevitably change a part of its meaning, so full content preservation cannot be reached. However, we should attempt to save the content as much as possible and adjust t^{δ} to the needs of this task.

Thus, the task of obtaining a text detoxification model with the best parameters set may be viewed as maximizing the probability $P(d^{tg}|d^{src}, s^{src}, s^{tg})$ given the three above-mentioned constraints based on parallel or non-parallel text corpora D^{src} and D^{tg}.

B CondBERT Ablation Study

Our CondBERT model was inspired by Wu et al. (2019a) and is similar to Wu et al. (2019b), but has some unique properties. We test their importance with the ablation study on the detoxification task.

Model	ACC	SIM	FL	J
Full model	0.91	0.73	0.75	0.50
Mask all toxic tokens	0.94	0.69	0.77	0.50
No similarity penalty	0.92	0.73	0.75	0.50
No multiword replacement	0.87	0.80	0.56	0.40
No toxicity penalty	0.57	0.79	0.82	0.33

Table 7: Results of the CondBERT ablation study on detoxification.

We use two heuristics for content preservation: not masking the toxic tokens and reranking replacement candidates with respect to their similarity to the original tokens. Removing any of these heuristics leads to lower content preservation and higher style accuracy, showing the inverse correlation of these properties (see Table 7). However, the **J** score for these models stays the same. On the other hand, turning off the possibility of filling a single mask with multiple words reduces the fluency and style accuracy, although obviously yields a better content preservation score, because the output sentence contains less new words. This affects the **J** score which is reduced for this model. Finally, the greatest impact on the **J** metric is caused by eliminating the toxicity penalty. The ACC is reduced dramatically, and although the other two metrics slightly grow compared to the full model, they cannot compensate for the low style accuracy.

C ParaGeDi Ablation Study

ParaGeDi model consist of a paraphraser, a language model trained with the generative-discriminative loss, and a style classifier for reranking hypotheses. In addition to that, we use a number of heuristics during inference. We conduct ablation study on the detoxification task to understand the usefulness of these components. We test the following variations of ParaGeDi:

- no discriminative loss ($\lambda = 0$),
- no generative loss ($\lambda = 1$),
- no upper bound ($u = \infty$),
- no discriminator ($w = 0$),

Model	ACC	SIM	FL	J
Full ParaGeDi	0.95	0.66	0.79	0.50
no reranking	0.87	0.70	0.80	0.50
beam of size 5	0.93	0.67	0.79	0.50
no discriminative loss	0.90	0.68	0.81	0.49
no smoothing	0.96	0.63	0.78	0.47
no beam search	0.88	0.66	0.70	0.41
no control of style strength	0.56	0.80	0.81	0.38
no generative loss	0.77	0.70	0.67	0.37
no upper bound	0.99	0.35	0.76	0.27

Table 8: Results of the ParaGeDi ablation study.

- no extra control of style strength ($w = 1$),
- no probability smoothing ($\alpha = 0$),
- no reranking,
- no beam search (beam size of 1).

In each configuration, all other parameters are fixed. The performance of models is given in Table 8.

Decreasing the number of beams leads to the deterioration of fluency and of style strength because of the smaller number options for the reranker to choose from. Removing the reranker leads to lower style strength with small gains in similarity or fluency. Turning off the smoothing of probabilities makes similarity and fluency degrade a little. Removing the upper bound on the discriminative correction leads to nearly 100% style transfer but to very low similarity of the generated sentences to the original ones, as the model starts hallucinating. Decreasing the w parameter reduces style accuracy but improves fluency and similarity, showing a clear trade-off between them.

The individual components of the loss are slightly less important for style than inference parameters. With only the discriminative loss the model is still able to successfully transform style in 77% of the cases, and the generative loss alone is able to change the style in 90% cases. The latter figure shows that the model equipped with style labels can discriminate between styles even if it was not explicitly trained to do that. On the other hand, the elimination of the generative loss results in a significant drop in fluency. Although the class-conditional LM in ParaGeDi is a GPT2 model which has already been trained for generation task, the lack of generation-based fine-tuning reduces the quality of the resulting text.

D Details of Mining the Parallel Corpus

Here we describe in more detail the process of mining of a detoxifying parallel paraphrase corpus. We use the ParaNMT dataset (Wieting and Gimpel, 2018) comprised of 50 million English sentences and their paraphrases back-translated from Czech. We filter the dataset keeping only the sentence pairs with moderate similarity (0.6 to 0.95) and similar length (with difference at most 40%), which is approximately 50% of the dataset. We compute the similarity as the cosine distance between the averaged BERT embeddings of all words in a sentence. After this similarity- and length-based filtering we score each sentence with a RoBERTa-based toxicity classifier from Section 5.1 and keep only the pairs with the difference in toxicity scores of at least 50%. Thus, we obtain 500,000 sentence pairs. Their examples are given in Table 9.

Manual inspection of a random sample of the selected pairs shows that around 10% of them are invalid paraphrases, 40% are in fact both toxic or both safe, and around 50% of them are valid detoxifying paraphrases. This suggests that with more rigorous filtering we can yield a corpus for detoxification of around 250,000 high-quality parallel sentences, which is larger than the majority of existing parallel style transfer datasets.

E Qualitative Analysis

Both automatic and manual joint scores show that our best models are halfway between useless and perfect. But the actual success rate is much less than half. We call a detoxified sentence "perfect" if all three annotators gave it the maximal scores for all three aspects. With this definition, only 20% of ParaGeDi sentences and 14% of CondBERT sentences are perfect, and only about 1.5% of Mask&Infill sentences are perfect.

As you can judge from Table 4, the main cause of imperfection for all models is distortion of meaning. Below we describe our manual investigation into the causes of this distortion.

In half of the cases, ParaGeDi conveys the meaning more or less adequately. Its mistakes include:

- replacement of toxic words by similarly looking less toxic words with different meaning (e.g. "whore" → "Who's who", "stop behaving like fascist thugs" → "Stop looking at fascism", "taxman massive cunt , only outdone by linuxcunt himself ." → "Taxman's massive cut, outdone by Linuxcune himself.").
- replacement of sentence meaning ("the election was yours to lose" → "the election is to be won", "this crap hole institute run by motherfuckers and

	Reference	Translation	sim	ld	tox_{ref}	tox_{trn}
0	If Alkar is flooding her with psychic waste, that explains the high level of neurotransmitters.	if Alkar floods her with her mental waste, it would explain the high levels of neurotransmitter.	0.78	0.01	0.01	0.98
1	Now you're getting nasty.	you're becoming disgusting.	0.75	0.07	0.06	0.99
2	Well, we could spare your life, for one.	well, we can spare your life.	0.92	0.27	0.21	0.98
3	Ah! Monkey, you've got to snap out of it.	monkey, you have to wake up.	0.66	0.31	0.05	0.99
4	I've got orders to put her down.	I have orders to kill her.	0.73	0.18	0.01	0.99
5	I'm not gonna have a child... ...with the same genetic disorder as me who's gonna die. L...	I'm not going to breed kids with a genetic disorder that makes them die.	0.70	0.20	0.95	0.04
6	They're all laughing at us, so we'll kick your ass.	they're laughing at us. We'll show you.	0.62	0.23	0.99	0.00
7	Maine was very short on black people back then.	there wasn't much black in Maine then.	0.72	0.19	0.96	0.15
8	Briggs, what the hell's happening?	Briggs, what the hell is going on?	0.92	0.00	0.16	0.84
9	Another one simply had no clue what to do, so whenever he met my brother he'd beat the crap out of him, and then say:	another simply didn't know what to do, so whenever he met my brother, he nearly beat the shit out of him.	0.88	0.10	0.05	0.93
10	I suppose you want me to buy you flowers and chocolates and whisper sweet nothings.	you'd probably want me to buy you some chocolates and flowers... and whispered some pretty rubbish.	0.80	0.16	0.00	0.98
11	So now their spirits are cursed, walking back roads, waterways, and if they find an unfaithful man, they kill him, and that man is never seen again.	their souls are cursed, they guard the paths, he says, and when they encounter an unfaithful man, he will be killed, and his body will never be found.	0.75	0.01	0.84	0.14
12	Freezing him.	I'll freeze him!	0.78	0.18	0.01	0.57
13	Come on, Cal, leave that shit alone.	come on, Cal, put it down.	0.66	0.27	0.99	0.00
14	So he's the Top dog.	he's the tallest son of a bitch.	0.61	0.36	0.00	0.99
15	I swore when I went out with Xander Harris... ...I'd rather die than datea fixer-upper again.	when I was dating Alex Harris, I swore I'd rather die than go out with a loser.	0.79	0.15	0.01	0.99
16	I'm famous, and you're done.	I'm famous, and you're dead.	0.82	0.00	0.00	0.98
17	To quote Jake Oppenheimer: I, who am about to die, must seem to them something "God-awful."...	to quote Jake and Oppenheimer: "I must die, I must feel like a terrible god."	0.70	0.18	0.00	0.68
18	"Could you please be quiet, Miss Lavish?" said Moist.	'could you keep your mouth shut, Miss Opulent? 'Said Moist.	0.81	0.10	0.00	0.76
19	Murder for hire.	murder to order.	0.70	0.00	0.07	0.96

Table 9: Examples of mined detoxifying paraphrases. Here **sim** is similarity between sentences, computed by (Wieting and Gimpel, 2018), **ld** is relative difference in length, $\textbf{tox}_{\text{ref}}$ and $\textbf{tox}_{\text{trn}}$ are toxicity scores calculated by our classifier.

bastards" → "a deloitte institute for mothers and their children")

- Avoiding the toxic or difficult part, for example "why we gotta have this miscegenation crap ?" → "Why do we need to have it?". In some cases, however, ParaGeDi masks or rephrases the toxic part of the message, while still preserving the general meaning, for example "start there first you idiot!" → "Let's start there first."

In general, ParaGeDi makes the impression of fantasising too much, because it often rewrites the whole sentence, and from time to time changes its structure significantly.

CondBERT, on the other hand, usually preserves the sentence structure, but often replaces words with inappropriate substitutes, often antonyms: "selfish" → "misunderstood", "racists" → "politicians", "cunt" → "nursemaid", "to troll and harass" → "to try out and help", "asskissers" → "honest people", "retarded" → "beautiful", "whore" → "sweetheart". Sometimes these replacements are more adequate, e.g. "old cock" → "old-fashioned stuff", "your attitude is shit" → "your attitude is completely wrong", "bitch i warned yall" → "i guess i warned you". But the proportion of sentences with the meaning preserved is smaller for CondBERT, being less than 40%.

The DLSM and Template-based DRG models often preserve the meaning by just preserving the toxic words, so their total success rate is low. The Retrieve-only DRG model almost never preserves the meaning. The Mask&Infill model seems to be overfitted: it often replaces toxic words with irrelevant non-toxic words (e.g. "crap" → "compelling") that the model apparently considers to be the "markers" of the non-toxic style. These properties make

the baselines unsuitable for the detoxification task without adaptation, and the CondBERT model is in fact such an adaptation.

Typical mistakes of both ParaGeDi and CondBERT can be attributed mostly to the insufficiency of semantic understanding: they often replace toxic words with semantically related words of different (often opposite) meaning, or simply with similarly looking words. We conjecture that with a paraphraser trained on a larger corpus (we have used only 2% of ParaNMT) or on more difficult examples would improve the ability of ParaGeDi to preserve meaning.

Generally, our models produce the impression of not being mature enough for fully automatic use with texts where meaning is important. However, they can be used to suggest detoxification options to human writers, or to detoxify the output of chit-chat bots where the cost of producing an inarticulate utterance is considerably less than the cost of producing a toxic one.

F Examples of Detoxification

Table 10 shows the examples of detoxification by our models ParaGeDi and CondBERT and the best competitor Mask&Infill.

Input	id have relatives signing up for aol finally in the mid to late 90s forwarding me stupid conservative conspiracy theory garbage .
ParaGeDi (ours)	I'd have relatives sign up for aol in the mid-to late 90s. and send me some unorthodox conspiracy theories.
CondBERT (ours)	id have relatives signing up for aol finally in the mid to late 90s forwarding me with some very conservative conspiracy theory stuff .
Mask&Infill	id have relatives signing up for aol finally in the mid to late 90s forwarding me to conservative conspiracy theory ideas .
Input	why dont you bung off . smoke a fag . and have a cuppa ?
ParaGeDi (ours)	Why don't you take a break. smoke a cigar. and have a cup of tea?
CondBERT (ours)	why dont you bung off . smoke a great deal . and have a cuppa ?
Mask&Infill	why dont you bung off . smoke a compelling compelling . and have a cuppa ?
Input	can someone please fix that crap ?
ParaGeDi (ours)	Can someone please do that?
CondBERT (ours)	can someone please fix that stuff ?
Mask&Infill	can someone please fix that compelling ?
Input	its a very stupid name !
ParaGeDi (ours)	It's a very strange name!
CondBERT (ours)	its a very . . . strange . . . name !
Mask&Infill	its a very compelling name !
Input	im a rookie pilot and would have never have done something this stupid .
ParaGeDi (ours)	I've been a pilot for a while. and I never would have done anything like
CondBERT (ours)	im a rookie pilot and would have never have done something this wrong .
Mask&Infill	im a rookie pilot and would have never have done something this compelling .
Input	cut the damn funding and lets move on to the next program .
ParaGeDi (ours)	Cut the funding and move on to the next program.
CondBERT (ours)	cut the amount of the funding and lets move on to the next program .
Mask&Infill	cut the compelling funding and lets move on to the next program .
Input	bitch i warned yall . didnt i ?
ParaGeDi (ours)	Bits I've warned a lot. right?
CondBERT (ours)	i guess i warned you . didnt i ?
Mask&Infill	compelling i warned yall . didnt i ?
Input	i was such an asshole .
ParaGeDi (ours)	I was an ad.
CondBERT (ours)	i was such an emotional mess .
Mask&Infill	i was such an compelling .
Input	your statements show me how ignorant you are .
ParaGeDi (ours)	Your statements show me that you'd be indifferent.
CondBERT (ours)	your statements show me how very wronged you are .
Mask&Infill	your statements show me how compelling you are .
Input	trump will see justin as a lightweight fool .
ParaGeDi (ours)	Trump will see justin as a light-hearted man.
CondBERT (ours)	trump will see justin as a lightweight world title contender .
Mask&Infill	trump will see justin as a lightweight compelling .
Input	this whole article is bullshit .
ParaGeDi (ours)	This article's not a good deal.
CondBERT (ours)	this whole article is considered incomplete .
Mask&Infill	this whole article is compelling .
Input	please dont revert my edits without good reasoning just makes you look stupid .
ParaGeDi (ours)	Please do not reverse my edits without a good reason.
CondBERT (ours)	please dont try to do my edits without good reasoning just makes you look pretty bad .
Mask&Infill	please dont revert my edits without good reasoning just makes you look compelling .
Input	do some research before you make a fool of yourself .
ParaGeDi (ours)	Do a bit of research.
CondBERT (ours)	do some research before you make a bigger picture out of my life .
Mask&Infill	do some research before you make a compelling of yourself .
Input	try not to make wikipedia look so stupid .
ParaGeDi (ours)	Try not to make Wikipedia seem like a bad idea.
CondBERT (ours)	try not to make wikipedia look so unsettling .
Mask&Infill	try not to make wikipedia look so compelling .
Input	i will make sure i revert any stupid edits you make from now on .
ParaGeDi (ours)	I'll be sure to correct any wrong edits that you make. from now on.
CondBERT (ours)	i will make sure i do not make any mistake about any edits you make from now on .
Mask&Infill	i will make sure i revert any compelling edits you make from now on .

Table 10: Examples of detoxification by the best-performing models considered in our study.