ParaDetox: Detoxification with Parallel Data

Varvara Logacheva[1*], **Daryna Dementieva**[1,3*], **Sergey Ustyantsev**[1], **Daniil Moskovskiy**[1], **David Dale**[1], **Irina Krotova**[2], **Nikita Semenov**[2], **and Alexander Panchenko**[1]
[1]Skolkovo Institute of Science and Technology, Russia
[2]Mobile TeleSystems (MTS), Russia
[3]Data and Web Science Group, University of Mannheim, Germany
{v.logacheva,daryna.dementieva,s.ustyantsev}@skoltech.ru,
{daniil.moskovskiy,d.dale,a.panchenko}@skoltech.ru,
{i.krotova@mts.ai, nikita.semenov}@mts.ru

Abstract

We present a novel pipeline for the collection of parallel data for the detoxification task. We collect non-toxic paraphrases for over 10,000 English toxic sentences. We also show that this pipeline can be used to distill a large existing corpus of paraphrases to get toxic-neutral sentence pairs. We release two parallel corpora which can be used for the training of detoxification models. To the best of our knowledge, these are the first parallel datasets for this task. We describe our pipeline in detail to make it fast to set up for a new language or domain, thus contributing to faster and easier development of new parallel resources.

We train several detoxification models on the collected data and compare them with several baselines and state-of-the-art unsupervised approaches. We conduct both automatic and manual evaluations. All models trained on parallel data outperform the state-of-the-art unsupervised models by a large margin. This suggests that our novel datasets can boost the performance of detoxification systems.

1 Introduction

Detection of toxicity (Zampieri et al., 2019) and other undesirable content, e.g. microaggressions (Breitfeller et al., 2019) or patronizing speech (Perez Almendros et al., 2020), is a popular topic of research in NLP. However, detection of harmful messages does not offer any proactive ways of fighting them (besides deletion). We suggest that such messages could be automatically rewritten to keep the useful content intact and eliminate toxicity.

The task of rewriting toxic messages (*detoxification*) has already been tackled by NLP researchers (Nogueira dos Santos et al., 2018; Tran et al., 2020). It is considered a variant of *style transfer* task, the task of rewriting a text saving the content and changing the style (*style* is defined as a characteristic of text such as sentiment, level of formality, or politeness, author profile (gender, political preferences), etc.). As a sequence-to-sequence task, style transfer can be performed with an encoder-decoder model trained on parallel data. However, there exist only a few parallel style transfer corpora (Carlson et al., 2018; Pryzant et al., 2020). Since they usually do not exist "naturally", they need to be written from scratch. This is an expensive and laborious process. Thus, such parallel datasets are extremely rare.

Jigsaw	so why would anyone believe this moron?
Para-phrase	so why would anyone believe this person? so why would anyone believe somebody like him?
Reddit	dude ham sandwich is the good sh*t .
Para-phrase	dude ham sandwich is the good thing The ham sandwich, buddy, is the bomb. Dude ham sandwich is good.
Twitter	now i feel like an a*s
Para-phrase	now i feel like worthless now i feel very bad now i feel bad

Table 1: Examples of detoxified sentences from the collected parallel corpus.

We aim at boosting the research in detoxification by collecting an English parallel corpus of toxic sentences and their non-toxic paraphrases. We suggest a new crowdsourcing pipeline for collecting parallel style transfer data. It does not employ experts, which makes the data collection faster and cheaper. In addition to generating the detoxified versions of texts, we consider a way to distill existing datasets of paraphrases for style-specific data. In particular, we find the pairs of toxic and non-toxic sentences in the paraNMT dataset (Wieting and Gimpel, 2018) of English paraphrases

[*] Equal contribution

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6804 - 6818
May 22-27, 2022 ©2022 Association for Computational Linguistics

and filter them using our crowdsourcing setup. The pipelines are described in detail to make them easy to replicate. Thus, we suggest that by reusing these pipelines the new parallel style transfer datasets can be collected in a fast and affordable way.

Finally, we validate the usefulness of our datasets by training detoxification models on them and comparing their performance with state-of-the-art methods. Models trained on parallel data significantly outperform other models in terms of automatic metrics and human evaluation.

The contributions of our work are three-fold:

- We suggest a novel pipeline for collection of parallel data for the detoxification task,
- We use the pipeline to collect the first parallel detoxification dataset **ParaDetox** (see Table 1 and Appendix A) and retrieve toxic-neutral pairs from **ParaNMT** corpus,[1]
- Using collected data we train supervised detoxification models that yield SOTA results.

2 Related Work

Style Transfer Datasets When collecting non-parallel style transfer corpora, style labels often already exist in the data (e.g. positive and negative reviews (Li et al., 2018)) or its source serves as a label (e.g. Twitter, academic texts, legal documents, etc.). Thus, data collection is reduced to fetching the texts from their sources, and the corpus size depends only on the available amount of text.

Conversely, parallel corpora are usually more difficult to get. There exist parallel style transfer datasets fetched from "naturally" available parallel sources: the Bible dataset (Carlson et al., 2018) features multiple translations of the Bible from different epochs, biased-to-neutral Wikipedia corpus (Pryzant et al., 2020) uses the information on article edits.

Besides these special cases, there exists a large style transfer dataset that was created from scratch. This is the GYAFC dataset (Rao and Tetreault, 2018) of informal sentences and their formal versions written by crowd workers and reviewed by experts. Since toxic-neutral pairs also do not occur in the wild, we follow this data collection setup with a notable difference – we replace expert validation of crowdsourced sentences with crowd validation and additionally optimize the cost.

Style Transfer and Detoxification The vast majority of style transfer models (including detoxification models) are trained on non-parallel data. They can perform pointwise corrections of style-marked words (Li et al., 2018; Wu et al., 2019; Malmi et al., 2020). Alternatively, some works train encoder-decoder models on non-parallel data and push decoder towards the target style using adversarial classifiers (Shen et al., 2017; Fu et al., 2018). As another way of fighting the lack of parallel data, researchers jointly train source-to-target and target-to-source style transfer models using reinforcement learning (Luo et al., 2019), amortized variational inference (He et al., 2020), or information from a style transfer classifier (Lee, 2020).

Detoxification is usually formulated as style transfer from toxic to neutral (non-toxic) style, so it uses non-parallel datasets labeled for toxicity and considers toxic and neutral sentences as two subcorpora. Laugier et al. (2021) use the Jigsaw datasets (Jigsaw, 2018, 2019, 2020) for training, Nogueira dos Santos et al. (2018) create their own toxicity-labelled datasets of sentences from Reddit and Twitter. Following them, we also fetch sentences for rewriting from these datasets.

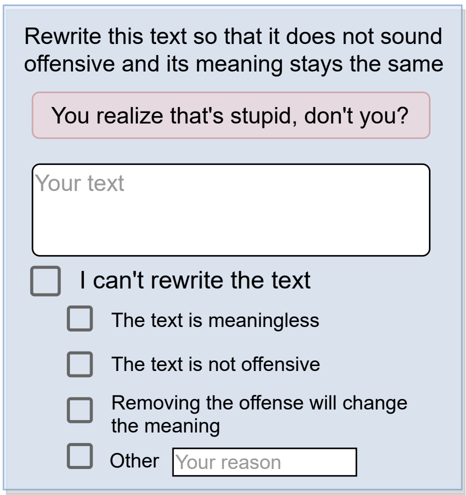
Works on detoxification often rely on style transfer models tested on other domains. Nogueira dos Santos et al. (2018) follow Shen et al. (2017) and Fu et al. (2018) and train an autoencoder with additional style classification and cycle-consistency losses. Laugier et al. (2021) perform a similar fine-tuning of T5 as a denoising autoencoder. Tran et al. (2020) apply pointwise corrections approach similar to that of Wu et al. (2019) and then improve the fluency of a text with a seq2seq model. Likewise, Dale et al. (2021) use a masked language model to perform pointwise edits of toxic sentences. They also suggest an alternative model which enhances a style-agnostic seq2seq model with style-informed language models which reweigh the seq2seq hypotheses with respect to the desired style.

When the parallel data is available, the majority of researchers use Machine Translation tools (Briakou et al., 2021) and pre-trained language models (Zhang et al., 2020) to perform style transfer. We follow this practice by fine-tuning BART model (Lewis et al., 2020) on our data.

3 Data Collection Pipeline

Our goal is to yield pairs of sentences that have the same meanings and are contrasted in terms of

[1] Our datasets and code of experiments is available online: https://github.com/skoltech-nlp/paradetox



Rewrite this text so that it does not sound offensive and its meaning stays the same

You realize that's stupid, don't you?

Your text

□ I can't rewrite the text

□ The text is meaningless

□ The text is not offensive

□ Removing the offense will change the meaning

□ Other Your reason

Figure 1: Interface of Task 1 (paraphrases generation).

offensiveness — one of the sentences is toxic and the other is neutral. We consider two scenarios: the manual rewriting of toxic sentences into neutral ones and the selection of toxic-neutral pairs from existing paraphrases. Unlike a similar work of Rao and Tetreault (2018), we hire crowd workers not only for the generation of paraphrases but also for their validation, which reduces both time and cost.

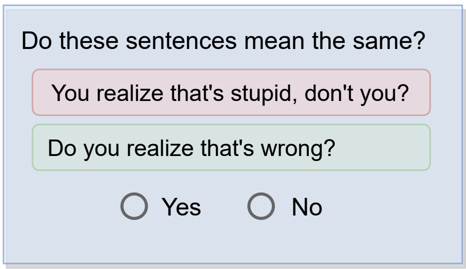
3.1 Crowdsourcing Tasks

We ask crowd workers to generate paraphrases and then evaluate them for content preservation and toxicity. Each task is implemented as a separate crowdsourcing project. We use the crowdsourcing platform Yandex.Toloka.[2]

Task 1: Generation of Paraphrases The first crowdsourcing task asks users to eliminate toxicity in a given sentence while keeping the content (see the task interface in Figure 1). However, it is not always possible. Some sentences cannot be detoxified, because they do not contain toxicity or because they are meaningless. Moreover, in some cases toxicity cannot be removed. Consider the examples:

- *Are you that dumb you can't figure it out?*
- *I've finally understood that wiki is nothing but a bunch of American racists.*

Not only the form but also the content of the messages are offensive, so trying to detoxify them would inevitably lead to a substantial change of sense. We prefer not to include such cases in the parallel dataset.



Do these sentences mean the same?

You realize that's stupid, don't you?

Do you realize that's wrong?

○ Yes ○ No

Figure 2: Interface of Task 2 (evaluation of content match).

If workers have to detoxify all inputs without a possibility to skip them, a large proportion of the generated paraphrases will be of low quality. Thus, we add the control "I can't rewrite the text" and optional controls to indicate the reasons.

Task 2: Content Preservation Check We show users the generated paraphrases along with their original variants and ask them to indicate if they have close meanings. Besides ensuring content preservation, this task implicitly filters out senseless outputs, because they do not keep the original content. The task interface is shown in Figure 2.

Task 3: Toxicity Check Finally, we check if the workers succeeded in removing toxicity. We ask users to indicate if the paraphrases contain any offense or swear words (see Figure 3).

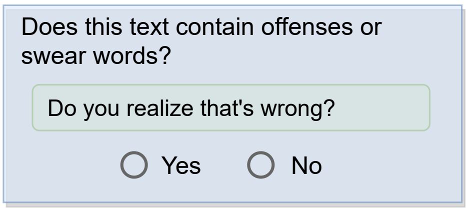
In addition to filtering out unsuitable paraphrases, we use Tasks 2 and 3 for paying for Task 1. We accept or reject the generated paraphrases based on the labels they get in Tasks 2 and 3.

3.2 Pipelines

Generation Pipeline To yield a parallel dataset, we first need to get toxic sentences for rewriting. We fetch them from corpora labeled for toxicity and additionally filter them with a toxicity classifier (described in Section 3.3). The overall data collection pipeline (see Figure 4) is as follows:

- Select toxic sentences for rewriting,
- Feed the sentences to **Task 1**,
- Feed the paraphrases generated in Task 1 to **Task 2**,
- Feed the paraphrases which passed Task 2 to **Task 3**,
- Pay for paraphrases from Task 1, if they passed checks in Task 2 and Task 3,
- Pay for "I can't rewrite" answers in Task 1 if two or more workers agreed on them.

[2] https://toloka.yandex.com



Does this text contain offenses or swear words?

Do you realize that's wrong?

○ Yes ○ No

Figure 3: Interface of Task 3 (evaluation of toxicity).

Retrieval Pipeline The generation pipeline can be used for cases when no parallel data is available. However, we suggest that a sufficiently large parallel corpus of paraphrases can contain pairs of sentences belonging to different styles, and it is possible to distill such corpus into a style transfer dataset. We check this hypothesis for the toxic and neutral styles on the ParaNMT dataset (Wieting and Gimpel, 2018).

We partially reuse the previously described setup. We do not need Task 1 since both toxic and neutral sentences are already available. However, we run Task 3 twice, because we need to check both parts of the pair for toxicity. Analogously to the generation pipeline, we use a toxicity classifier to pre-select pairs of sentences where one sentence is toxic and the other one is neutral. The parallel data retrieval pipeline is shown in Figure 5. It is simpler because Tasks 2 and 3 do not serve for paying for the generated paraphrases and are only used for data filtering. The pipeline is as follows:

- Select a pair of sentences (toxic and non-toxic) from the parallel data,
- Feed the toxic sentence candidate to **Task 3** to make sure it is toxic,
- Feed the neutral sentence candidate to **Task 3** to make sure it is non-toxic,
- Feed both sentences to **Task 2** to check if their content matches.

3.3 Crowdsourcing Settings

Preprocessing To pre-select toxic sentences, we need a toxicity classifier. We fine-tune a RoBERTa model (Liu et al., 2019)[3] on half of the three merged Jigsaw datasets (Jigsaw, 2018, 2019, 2020) (1 million sentences) and get a classifier which yields the F_1-score of 0.76 on the Jigsaw test set (Jigsaw, 2018). We consider a sentence toxic if the classifier confidence is above 0.8. To make the sentences easier for reading and rewriting, we choose the ones consisting of 5 to 20 tokens. For the retrieval pipeline, we also select parallel sentences with the cosine similarity of embeddings between 0.65 and 0.8. The similarity scores were provided as a part of ParaNMT dataset, the embeddings come from the PARAGRAM-PHRASE model (Wieting et al., 2016). Based on a manual validation, sentences with lower similarity are often not exact paraphrases, and too-similar sentences are either both toxic or both non-toxic.

Quality Control To perform paid tasks, users need to pass *training* and *exam* sets of tasks. Each of them has a corresponding *skill* – the percentage of correct answers. It is assigned to a user upon completing training or exam and serves for filtering out low-performing users. Besides that, users are occasionally given control questions during labeling. They serve for computing the *labeling skill* which can be used for banning low-performing and rewarding well-performing workers. The overall training and control pipeline is shown in Figure 6. It is used in Tasks 2 and 3.

In Task 1 we perform different quality control. We ban users who submit answers which are: (i) a copy of the input, (ii) too short (< 3 tokens) or too long (more than doubled original length), (iii) contain too many rare words or non-words. The latter condition is checked as follows. We compute the ratio of the number of whitespace-separated tokens and the number of tokens identified by the BPE tokeniser (Sennrich et al., 2016).[4] The rationale behind this check is that the BPE tokenizer tends to divide rare words into multiple tokens. If the number of BPE tokens in a sentence is two times more than the number of regular tokens, it might indicate the presence of non-words. We filter out these answers and ban users who produce them.

In addition to that, we ban malicious workers using built-in Yandex.Toloka tools: (i) **captcha**, (ii) number of **skipped questions** — we ban users who skip 10 task pages in a row, and (iii) **task completion time** — we ban those who accomplish tasks too fast (this usually means that they choose a random answer without reading).

Payment In Yandex.Toloka, a worker is paid for a page that can have multiple tasks (the number is set by customer). In Task 1, a page contains 5 tasks and costs \$0.02. In Tasks 2 and 3, we pay \$0.02

[3] https://huggingface.co/roberta-large

[4] We use the tokenizer of the BERT base uncased model (https://huggingface.co/bert-base-uncased)

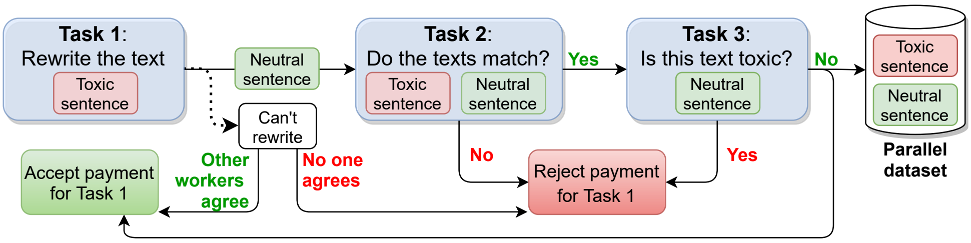


Figure 4: The pipeline of crowdsourcing for generation of detoxifying paraphrases.

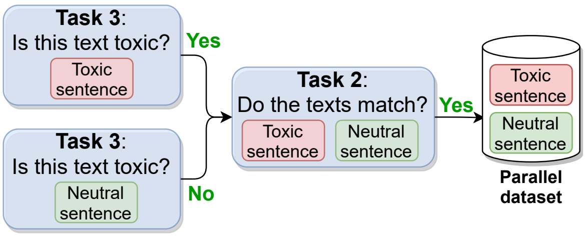


Figure 5: The retrieval pipeline.

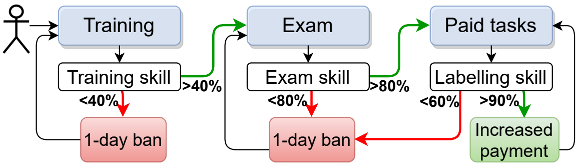


Figure 6: Training and quality control pipeline for Tasks 2 and 3.

and $0.01, respectively, for 12 tasks. In addition to that, in these tasks, we use skill-based payment. If a worker has the *labeling skill* of above 90%, the payment is increased to $0.03 (Task 2) and $0.02 (Task 3).

Tasks 2 and 3 are paid instantly, whereas in Task 1 we check the paraphrases before paying. If a worker indicated that a sentence cannot be paraphrased, we pay for this answer only if at least one other worker agreed with that. If a worker typed in a paraphrase, we send it to Tasks 2 and 3 and pay only for the ones approved by both tasks. The payment procedure is shown in Figure 4.

Postprocessing To ensure the correctness of labeling, we ask several workers to label each example. In Task 1, this gives us multiple paraphrases and also verifies the "I can't rewrite" answers. For Tasks 2 and 3, we compute the final label using the Dawid-Skene aggregation method (Dawid and Skene, 1979) which defines the true label iteratively giving more weight to the answers of workers who agree with other workers more often. The number of people to label an example ranges from 3 to 5 depending on the workers' agreement.

Dawid-Skene aggregation returns the final label and its confidence. To improve the quality of the data, we accept only labels with the confidence of over 90% and do not include the rest in the final data.

3.4 The Pipeline Scalability

The Yandex.Toloka platform has an interface in English and workers from a large number of countries. Workers can be filtered by their location and asked to pass built-in language tests (available for many languages) to ensure the knowledge of a particular language. This enables the use of Toloka for the creation of NLP resources in many languages.

In our work, crowd workers manually rephrase sentences from non-parallel datasets. The pipeline does not require any specific data format and can be applied to any text. The only prerequisites are to define the source and target styles and to formulate the task of transferring between them. Thus, we believe that the pipeline is suitable for creating parallel datasets for any other style transfer tasks, at least those which have non-parallel datasets and clear definitions of style (positive ↔ negative, complex ↔ simple, impolite ↔ polite, etc.).

We should admit that our pipeline suggests the availability of (non-parallel) datasets in the chosen styles or at least publicly available sources of such data (e.g. social networks, question answering platforms). However, this is also a prerequisite for any style transfer model trained on non-parallel data. Therefore, any work on style transfer suggests that there exists enough data in the chosen style pair and language. This should not be considered a specific limitation of the pipeline.

4 Data Analysis

We collected **ParaDetox** – a parallel detoxification dataset with 1–3 paraphrases for over 12,000 toxic sentences. We also manually filtered **ParaNMT** dataset and get 1,400 toxic-neutral pairs.

4.1 ParaDetox: Generated Paraphrases

We fetched toxic sentences from three sources: Jigsaw dataset of toxic sentences (Jigsaw, 2018), Reddit and Twitter datasets used by Nogueira dos Santos et al. (2018). We selected 7,000 toxic sentences from each source and gave each of the sentences for paraphrasing to 3 workers. We get paraphrases for 12,610 toxic sentences (on average 1.66 paraphrases per sentence), 20,437 paraphrases total. Running 1,000 input sentences through the pipeline costs $41.2, and the cost of one output sample is $0.07. The overall cost of the dataset is $811.55. We give them examples of sentences in Appendix A. In addition to that, we provide some samples which could not be detoxified in Appendix C. The statistics of the paraphrases written by crowd workers are presented in Table 2.

The distribution of sentences from different datasets in the final data is not equal. Jigsaw turned out to be the most difficult to paraphrase. Fewer sentences from it are successfully paraphrased, making it the most expensive part of the collected corpus ($0.08 per sample). Figure 7 shows that the number of untransferable sentences in the Jigsaw dataset is larger than that of other corpora.

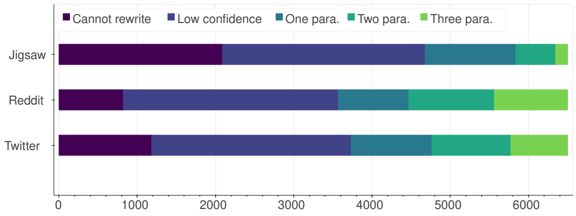


Figure 7: Number of paraphrases per input.

Out of all crowdsourced paraphrases, only a small part was of high quality. We plot the percentage of paraphrases which were filtered out by content and toxicity checks in Figure 8. It also corroborates the difficulty of the Jigsaw dataset. While the overall number of generated paraphrases was slightly higher for it, much more of them were discarded.

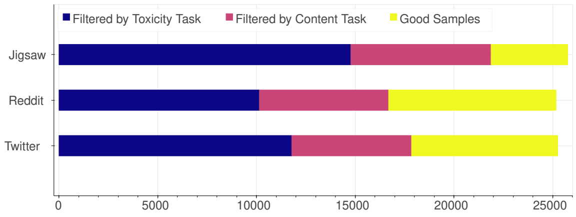


Figure 8: Data filtering output.

4.2 Analysis of Edits

Although we did not give any special instructions to workers about editing, they often followed the minimal editing principle, making 1.36 changes per sentence on average. A change is deletion, insertion, or rewriting of a word or multiple adjacent words. Many of the changes are supposedly deletions because the average sentence length drops from 12.1 to 10.4 words after editing.

The nature of editing differs for the three datasets. We compute the percentage of edits which consisted of removing the most common swear words or replacing them with neutral words. We first define the differences between the original and transformed string with the `difflib` Python library and then compute the percentage of differences that consist in editing swear words and other (non-offensive) words. We use a small manually compiled list of swear words which includes words *f*ck*, *sh*t*, *a*s*, *b*tch*, *d*mn* and their variants. Table 3 shows that the deletion or replacements of the most common swearing constituted a large part of all edits for Reddit and Twitter datasets (22% and 30%), while for Jigsaw it was only 3%.

Another surprisingly common type of editing is the normalization of sentences. The users often fixed casing, punctuation, typos (e.g. *dont* → *don't*, *there's* → *there is*). They also tended to replace colloquial phrases with more formal and standard language. Finally, some users overcorrected the sentences. For example, they replaced neutral words such as *dead*, *murder*, *penis* with euphemisms. This tendency indicates that workers consider any sensitive topic to be inappropriate content and try to avoid it as much as possible.

4.3 ParaNMT: Existing Paraphrases

Our automatic filtering of ParaNMT for content yields 500,000 potentially detoxifying sentence pairs, which is 1% of the corpus. We then sample 6,000 random pairs from this list and ask workers to evaluate them for toxicity and content preservation. This leaves with 1,393 sentences, meaning that around 23% of the pre-selected sentence pairs were approved (for ParaDetox we get paraphrases for 61% input sentences). Thus, although the cost per 1,000 inputs is much lower than that of generating the paraphrases, the cost per output sample is the same as that of generated paraphrases.

ParaNMT dataset is different from ParaDetox. First, each sentence has only one paraphrase. These

Source Dataset	Input Samples	Unique Inputs Paraphrased	Paraphrases per Inputs	Paraphrases Total	Edits per sample	Sample length (tokens) toxic/neutral	Cost per 1,000 inputs	Cost per unique sample
ParaDetox (Generated paraphrases)								
Jigsaw	7,000	3,054	1.34	4,082	1.32	12.0 / 10.4	$36.65	$0.08
Reddit	7,000	4,947	1.75	8,681	1.34	12.4 / 10.7	$47.77	$0.06
Twitter	7,000	4,609	1.55	7,674	1.4	11.9 / 10.1	$42.30	$0.06
Total	21,000	12,610	1.62	20,437	1.36	12.1 / 10.4	$41.18	$0.07
ParaNMT (Existing paraphrases)								
ParaNMT	6,000	–	1	1,393	2.54	12.7 / 11.7	$17.40	$0.08

Table 2: Statistics of the crowdsourcing experiments and final datasets.

	Swear words		Other phrases		
Dataset	Del	Rep	Del	Rep	Ins
Jigsaw	2.3%	0.6%	30%	60%	6.8%
Reddit	19%	9.1%	26%	41%	5.7%
Twitter	15%	7.1%	23%	47%	8.2%
ParaNMT	1.6%	1.2%	19%	64%	14%

Table 3: Percentage of common swear words (f*ck, sh*t, a*s and their common variants) and other words **Del**eted, **Rep**laced, or **Ins**erted by crowd workers.

paraphrases were not gained via manual editing but via a chain of translation models. Thus, neutral sentences are less similar to the toxic sentences, and the edits are more diverse, which makes it more similar to Jigsaw dataset (see Table 3).

5 Evaluation

To evaluate the collected corpora, we use them to train several supervised detoxification models. We separate the ParaDetox dataset into training and test parts (11,939 and 671 sentence pairs, respectively). The test sentences have one reference per sentence. We manually validate the test set to exclude the appearance of non-detoxifiable sentences or sentences which stayed toxic after rewriting (we need to verify that since the corpus was generated via crowdsourcing only). We do not use the test set neither for training nor for parameter selection of the models.

5.1 Models

We fine-tune a Transformer-based generation model BART (Lewis et al., 2020)[5] on our data. We test BART trained on the following datasets:

- **ParaDetox** – our full crowdsourced dataset.
- **ParaDetox-unique** – a subset of ParaDetox where each toxic sentence has only one paraphrase (selected randomly).
- **ParaDetox-1000** – 1,000 samples from the crowdsourced dataset (distributed evenly across data sources, each toxic sample has multiple non-toxic variants).
- **ParaNMT** – filtered ParaNMT corpus, **auto** stands for automatically filtered 500,000 samples, **manual** are 1,393 manually selected sentence pairs.

We train BART for 10,000 epochs with the learning rate of 3e-5 and the number of gradient accumulation steps set to 1. The other parameters are set to their default values.

We also compare our models to other style transfer approaches:

- **Duplicate** (baseline) – copy of the input,
- **Delete** (baseline) – deletion of swear words,
- **BART-zero-shot** (baseline) – BART model with no additional training.
- **Mask&Infill** (Wu et al., 2019) – BERT-based pointwise editing model,
- Delete-Retrieve-Generate models (Li et al., 2018): **DRG-Template** (replacement of toxic words with similar neutral words) and **DRG-Retrieve** (retrieval of non-toxic sentences with the similar sense) varieties.
- **DLSM** (He et al., 2020) encoder-decoder model that uses amortised variational inference,
- **SST** (Lee, 2020) – encoder-decoder model with the cross-entropy of a pretrained style classifier as an additional discriminative loss.
- **CondBERT** (Dale et al., 2021) – BERT-based model with extra style and content control,

[5] We use model https://huggingface.co/facebook/bart-base

- **ParaGeDi** (Dale et al., 2021) – a model which enhances a paraphraser with style-informed LMs which re-weigh its output.

5.2 Metrics

We compute the BLEU score on the test set. In addition to that, we perform automatic reference-free evaluation which is used in many style transfer works. Namely, we evaluate:

- *Style accuracy* (**STA**) – percentage of non-toxic outputs identified by a style classifier. We use a classifier from Section 3.3 trained on a different half of Jigsaw data.
- *Content preservation* (**SIM**) – cosine similarity between the embeddings of the original text and the output computed with the model of Wieting et al. (2019). This model is trained on paraphrase pairs extracted from ParaNMT corpus. The model's training objective is to yield embeddings such that the similarity of embeddings of paraphrases is higher than the similarity between sentences that are not paraphrases.
- *Fluency* (**FL**) – percentage of fluent sentences identified by a RoBERTa-based classifier of linguistic acceptability trained on the CoLA dataset (Warstadt et al., 2019).

	BLEU	STA	SIM	FL	J
Human reference	100.0	0.96	0.77	0.88	0.66
Baselines and SOTA (unsupervised)					
Delete	61.24	0.81	0.93	0.64	0.46
Duplicate	53.86	0.02	1.0	0.91	0.02
DRG-Template	53.86	0.90	0.82	0.69	0.51
BART-zero-shot	53.64	0.01	**0.99**	0.92	0.01
Mask&Infill	52.47	0.91	0.82	0.63	0.48
CondBERT	42.45	**0.98**	0.77	0.82	0.62
SST	30.20	0.86	0.57	0.19	0.10
ParaGeDi	25.39	**0.99**	0.71	0.88	0.62
DLSM	21.13	0.76	0.76	0.52	0.25
DRG-Retrieve	4.74	0.97	0.36	0.86	0.31
BART on parallel data (supervised) – *our models*					
ParaDetox	**64.53**	0.89	0.86	0.89	**0.68**
ParaDetox-unique	**64.58**	0.87	0.87	0.88	0.65
ParaDetox-1000	63.26	0.83	0.86	0.90	0.62
ParaNMT-man	46.58	0.76	0.81	**0.93**	0.55
ParaNMT-auto	43.30	0.62	0.85	**0.94**	0.48

Table 4: Automatic evaluation of detoxification models. Numbers **in bold** indicate the best results. Rows in gray indicate the baselines.

We compute the final joint metric (**J**) as the multiplication of the three individual metrics.

Since the automatic evaluation can be unreliable, we evaluate some models manually. We randomly select 200 sentences from the test set and ask assessors to evaluate them along the same three parameters: style accuracy (**STA**$_m$), content preservation (**SIM**$_m$), and fluency (**FL**$_m$). All parameters can take values of 1 (good) and 0 (bad). We also report the joint metric **J**$_m$ which is the percentage of sentences whose **STA**$_m$, **SIM**$_m$, and **FL**$_m$ are 1.

The evaluation was conducted by 6 NLP researchers with a good command of English. Each sample was evaluated by 3 assessors. The inter-annotator agreement (Krippendorff's α) reaches 0.64 (**STA**$_m$), 0.67 (**SIM**$_m$), and 0.68 (**FL**$_m$).

5.3 Results

Automatic Evaluation Table 4 shows the automatic scores of all tested models. Our BART models trained on ParaDetox outperform other systems in terms of BLEU and J. The much lower scores of BART-zero-shot confirm that this success is due to fine-tuning and not the innate ability of BART. The majority of unsupervised SOTA approaches are not only worse than BART but also perform below the "change nothing" baseline. The closest competitor of our models is the Delete model. This can be explained by the fact that crowd workers often only remove or replaced swear words which is what the Delete model does.

When comparing models trained on supervised data, we can see that BART does not benefit from multiple detoxifications per sentence, its performance is the same when trained on ParaDetox and ParaDetox-unique. On the other hand, manual filtering of ParaNMT is beneficial, it increases the quality of BART trained on it, although the number of training sentences drops from 500,000 to 1,400.

We also check which amount of data is sufficient for a high detoxification quality. We train the BART model on subsets of ParaDetox of different sizes. Figure 9 and the performance of ParaDetox-1000 model (Table 4) show that 1,000 training samples is enough to get a good detoxification. While SIM and FL are already high for vanilla BART (see BART-zero-shot model), STA can be improved with only a few parallel examples. This suggests that style transfer does not need large parallel corpora, making our pipeline more useful for other style transfer tasks. However, this is the result of the automatic evaluation, which as we show below is not always reliable. It needs extra investigation.

Original	economies of venezuela, iraq, etc still shit .	f*ck you, i wont do what you tell me.	your types of examples are idiotic.
Delete	economies of venezuela , iraq, etc still .	you, i wont do what you tell me.	your types of examples are.
CondBERT	economies of venezuela , iraq , etc still exist today.	unless i tell you, i wont do what you tell me.	your types of examples are very interesting.
ParaGeDi	economies of venezuela, iraq, etc still intact.	Fick, I'll do what you say.	Your types of examples are weird.
BART-ParaDetox	**economies of venezuela, iraq etc are still bad.**	**I won't do what you tell me.**	**Your types of examples are not good**.

Table 5: Examples of detoxifications by different models. Bad answers are shown in red, best answers **in bold**.

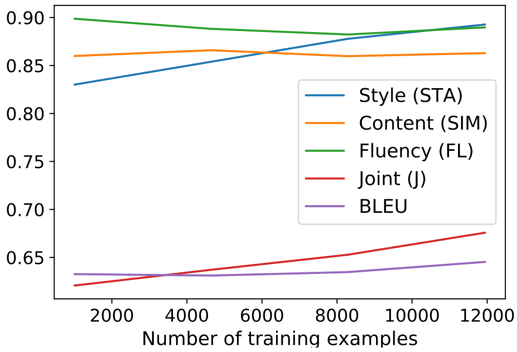


Figure 9: Scores of BART models trained on parallel data subsets of different sizes.

Table 5 shows examples of different models output. Delete performs deterministic operations which can return disfluent text. CondBERT has to insert something instead of a toxic word, which is not always a good strategy. ParaGeDi generates sentences from scratch, which sometimes results in a distorted sense. BART trained on parallel data is usually free of these drawbacks. More examples of outputs are available in Appendix B.

	STA$_m$	SIM$_m$	FL$_m$	J$_m$
Delete	0.785	0.445	0.365	0.21
CondBERT	**0.935**	0.250	0.615	0.15
ParaGeDi	**0.930**	0.415	0.870	0.37
BART-ParaDetox	0.830	**0.925**	**0.960**	**0.76**
BART-ParaNMT-man	0.750	0.705	**0.960**	0.50

Table 6: Manual evaluation of detoxification models. Numbers **in bold** indicate the best results (with the statistical significance $\alpha = 0.01$).

Manual Evaluation Manual evaluation (Table 6) confirms the usefulness of parallel data. BARTs trained on parallel data outperform other competitors, even if the size of this data is small. However, manual and automatic evaluations do not always match. Here, the well-performing Delete model gets the lowest score.

Overall, assessors agree with automatic metrics only in terms of fluency, their Spearman correlation r is 0.89. The manual style accuracy and content preservation are only moderately correlated with their automatic counterparts leaving space for further improvements. J and J$_m$ almost do not correlate. Besides that, BLEU correlates only with content preservation score and is moderately inversely correlated with the style accuracy. Thus, BLEU measures only the degree of content preservation and cannot replace other metrics.

6 Conclusions

We present ParaDetox – an English parallel corpus for the detoxification task. It contains almost 12,000 user-generated toxic sentences manually rewritten by crowd workers. To the best of our knowledge, this is the first parallel detoxification dataset. We present a novel data collection pipeline and show that parallel data can be generated using only crowdsourcing. We also adopt this pipeline to the style-based distillation of paraphrase corpus.

We confirm the usefulness of our datasets by training sequence-to-sequence models on them. The experiments show that the use of parallel data yields models which significantly outperform style transfer models trained on non-parallel data. Besides that, we confirm that filtering the noisy parallel data can lead to considerable improvement.

We see that it is enough to get 1,000 parallel sentences to perform detoxification with high quality. This suggests that our pipeline can be successfully applied to create useful parallel resources for style transfer even in cases of limited finance or lack of crowd workers because the cost of generating 1,000 examples is very low.

Finally, we investigate the relationship between metrics and find that automatic evaluation does not always match the manual judgments and reference-based BLEU cannot replace human evaluation, because it measures content preservation.

Acknowledgements

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

Ethical Considerations

The research on toxicity raises some ethical issues. In terms of our work, the parallel corpus we created can indeed be used in the reverse direction, i.e. to "toxify" sentences. However, although we did not thoroughly evaluate the quality of such toxification, our intuition is that it would not be high enough to make the corrupted sentences look natural. The reason is that the toxic part of our corpus consists of real toxic sentences fetched on the Internet, whereas their non-toxic counterparts are "translations" performed by crowd workers. We suggest that they obey the common regularities observed for *translationese* (texts manually translated from their original language into a different one): they differ from regular texts in terms of vocabulary (Koppel and Ordan, 2011) and syntax (Lembersky et al., 2011). The manually detoxified texts are different from the original non-toxic texts written by Internet users from scratch. While they are still recognized by human assessors as plausible sentences, we suggest that a sequence-to-sequence model trained to get translationese as input would not be as successful in transforming real texts (as it was shown for machine translation models (Freitag et al., 2019)).

Thus, although our corpus can be used in the reverse direction, it is not symmetric, which makes it less efficient as training datasets for "toxifiers". However, we should emphasize that these statements are our hypotheses and should be further investigated. Finally, we argue that the risk of using our corpus for toxification is perhaps not game-changing, as simpler approaches based on patterns (e.g. including a set of predefined obscene fragments into neutral texts) can serve the same purpose relatively well.

References

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. 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.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society Open Science*, 5.

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

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

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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.

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.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

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*, pages 195–204, Dublin, Ireland. Association for Computational Linguistics.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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.

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.

Cicero 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 (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. 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.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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.

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*, pages 2107–2114, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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 Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. 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 (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5271–5277.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yi Zhang, Tao Ge, and Xu Sun. 2020. 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.

A ParaDetox Samples

Table 7 shows the examples of manually detoxified parallel sentences from the ParaDetox corpus.

Original	my computer is broken and my phone too!! wtf is this devil sh*t???
Paraphrases	My computer is broken and my phone too! So disappointed! My computer is broken and my phone too, what is this? Both my computer and phone are broken.
Original	sh*t is crazy around here.
Paraphrases	It is crazy around here. Stuff is crazy around here. Something is crazy around here.
Original	delete the page and shut up
Paraphrases	Delete the page and stay silent. please delete the page delete the page
Original	massive and sustained public pressure is the only way to get these b*stards to act.
Paraphrases	Massive and sustained public pressure is the only way to get them to act. massive and sustained preasure is the only way to get these people to act
Original	f*ck you taking credit for some sh*t i wanted to do
Paraphrases	You are taking credit for something I wanted to do You're taking credit fro something i wanted to do.
Original	you gotta admit that was f*ckin hilarious though!
Paraphrases	you got to admit that was very hilarious though! you gotta admit that was hilarious though!
Original	as an american who thought it was a d*ck joke , thanks.
Paraphrases	as an american who thought it was a joke, thanks As an American who thought it was a bad joke, thanks
Original	for whatever reason , your comment just blew my f*cking mind.
Paraphrases	For whatever reason, your comment just blew my mind. for whatever reason, your comment just amazes me.
Original	what exactly is your f*cking problem here?
Paraphrases	What exactly is your problem here?
Original	who the f*ck are you gona call when that happens.
Paraphrases	Who are you gonna call when that happens?
Original	some idiots no longer believe in national sovereignty.
Paraphrases	Some people no longer believe in national sovereignty
Original	i was f*ckin bored as sh*t
Paraphrases	I was bored

Table 7: Examples of detoxified sentences from the collected parallel corpus.

B Outputs of Detoxification Models

Table 8 contains the outputs of four well-performing detoxification models: Delete model which deterministically removes swear words, unsupervised ParaGeDi and CondBERT models (Dale et al., 2021), and BART model fine-tuned on our parallel ParaDetox dataset.

Original	at least one party gives a cr*p about humans after birth.
Delete	at least one party gives a about humans after birth.
CondBERT	at least one party gives a whole story about humans after birth.
ParaGeDi	At the same time, at least one party doesn't believe in human life after birth
BART-ParaDetox	**at least one party cares about humans after birth.**
Original	gosh, discrimination is such a b*tch!
Delete	gosh, discrimination is such a!
CondBERT	gosh, discrimination is such a thing!
ParaGeDi	Gosh, discrimination is a bit of a bad thing!
BART-ParaDetox	**Discrimination is bad!**
Original	he ll f*ck that up too.
Delete	he ll that up too.
CondBERT	he ll ' ve picked that up too.
ParaGeDi	He'll get that up.
BART-ParaDetox	**He will mess that up too.**
Original	whats the title of this report, no sh*t?
Delete	whats the title of this report, no?
CondBERT	whats the title of this report , no stuff?
ParaGeDi	What's the name of the report, no?
BART-ParaDetox	**whats the title of this report?**
Original	because its f*cking reposted so much
Delete	**because its reposted so much**
CondBERT	because its parents have been reposted so much
ParaGeDi	Because it'd been so much re-enacted.
BART-ParaDetox	**Because it reposted so much.**
Original	to suggest otherwise is stupidly inflammatory.
Delete	to suggest otherwise is stupidly inflammatory.
CondBERT	to suggest otherwise is not necessarily inflammatory.
ParaGeDi	Saying otherwise would be a bad thing.
BART-ParaDetox	**To suggest otherwise is inappropriate.**
Original	he hasn t done sh*t for palestine.
Delete	he hasn t done for palestine.
CondBERT	he hasn t done much else for palestine.
ParaGeDi	He hasn't done it for palestine.
BART-ParaDetox	**He hasn't done anything for palestine.**

Table 8: Examples of detoxifications by different models. Bad answers are shown in red, the best answers **in bold**.

C Non-detoxifiable Samples

Table 9 gives examples of sentences which could not be detoxified by any of workers who tried to do that. We group the samples by the reasons of the inability to detoxify that the workers indicated.

Unclear content
she is! f*ck's sake stross! *fights against you*
yourwhy bring in the red herring of extramarital sex, getting desperate aj?
wikibofh f*ck u u f*ckin suck argument is fine u nerd now unblock me.
you have to be a complete idiot to be supporting the gop ticket this go around.
wikipedia review and tell the world why you not a grand d*ckhead
trump will make a fool out of little potato
dont dumb as in dead dummy.
justin and his rat pack have an agenda.
trump was an embarrassment to the u
that guy knew how to kill commies like nobodys business!
some stinks cannot be unstunk.
Toxic content
f*ck you you mother f*cking a*s cracker dildo licking piece of f*cking a*s sh*t!
if by christian you mean hypocrites.
lying anti american muslim mouthpiece.
they simply are too stupid to do otherwise.
xman is either an idiot or a psychopath or maybe both.
youre the insignificant f*ck mother f*cker!
hello, you sexy little lovehole!
obviously feminists cant be good mothers or wives.
we are country of losers and abusers.
are you a troublemaker or just plain stupid or both?
as anyone with half a brain would do.
No toxicity
this is all coming out of our darned pockets, and i am not confident!
you people complete lack morals.
this idea seems to be a rather ignorant one.
youre implying, therefore, that women ought to stay away from all black men.
blaming everyone else for the hole that you dug is pathetic.
killing the innocent nearly born should be the very last choice.
ignorant to me means without knowledge.
how can students of colour be expected to learn in such a toxic environment of white supremacy?
the problem is that their management is so ridiculously incompetent.
trump will keep on committing political suicide.
making stupid remarks is useless, do some research and then make a comment.

Table 9: Examples of sentences which could not be detoxified for different reasons.