

Crowdsourcing of Parallel Corpora: the Case of Style Transfer for Detoxification

Daryna Dementieva¹, Sergey Ustyantsev¹, David Dale¹, Olga Kozlova²,
Nikita Semenov², Alexander Panchenko¹ and Varvara Logacheva¹

¹Skolkovo Institute of Science and Technology, Moscow, Russia

²Mobile TeleSystems (MTS), Moscow, Russia

Abstract

One of the ways to fighting toxicity online is to automatically rewrite toxic messages. This is a sequence-to-sequence task, and the easiest way of solving it is to train an encoder-decoder model on a set of parallel sentences (pairs of sentences with the same meaning, where one is offensive and the other is not). However, such data does not exist, making researchers resort to non-parallel corpora. We close this gap by suggesting a crowdsourcing scenario for creating a parallel dataset of detoxifying paraphrases. In our first experiments, we collect paraphrases for 1,200 toxic sentences. We describe and analyse the crowdsourcing setup and the resulting corpus.

Keywords

toxicity, dataset, crowdsourcing, parallel data

1. Introduction

Toxicity is a major problem on the Internet. There are multiple strategies for fighting it. Detection of toxicity, being a popular area of research in NLP [1, 2, 3], is a necessary first step for any toxicity removal strategy. There exist multiple methods for toxicity detection [4, 5] and datasets [6, 7, 8] which can be used as training data for toxicity detection models.

Toxic messages can differ in terms of their content. For some of them, the offence is the only content, and their only goal is to insult the reader. On the other hand, other messages classified as toxic can contain a non-toxic and useful content that is only expressed in a toxic way. If we manage to rewrite such messages to eliminate the toxicity and save the content, we will be able to keep a constructive and peaceful discussion and prevent it from decaying to a row of insults.


The task of detoxification of text has already been tackled by several researchers [9, 10, 11]. This task is usually formulated as a style transfer task [12] — a task of rewriting a text with keeping the original content as much as possible and changing a particular attribute (author profile, text sentiment, degree of complexity, etc.). This attribute is referred to as style. In the case of detoxification, the task is to transfer text from toxic to neutral (non-toxic) style. It is similar to the tasks of rewriting texts with a varying degree of formality [13, 14] or politeness [15, 16].

VLDB 2021 Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, August 20, 2021, Copenhagen, Denmark

✉ daryna.dementieva@skoltech.ru (D. Dementieva); s.ustyantsev@skoltech.ru (S. Ustyantsev);
d.dale@skoltech.ru (D. Dale); oskozlo9@mts.ru (O. Kozlova); nikita.semenov@mts.ru (N. Semenov);
a.panchenko@skoltech.ru (A. Panchenko); v.logacheva@skoltech.ru (V. Logacheva)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

As with many other sequence-to-sequence tasks (machine translation, text summarization, paraphrasing), style transfer can be successfully solved using encoder-decoder models trained on parallel data [14, 17]. This is the easiest training setup because it contains a correct answer towards which a model can be optimised. However, for the majority of style transfer tasks, the parallel data is unavailable [12]. Therefore, style transfer models are usually trained on non-parallel datasets. If parallel data was available, they could perform better with smaller training corpora. The experiments show that training even on an extremely small parallel dataset can yield a well-performing style transfer model [18].

The goal of our work is to close this gap for one of the directions of text style transfer, namely detoxification. We collect a parallel dataset via crowdsourcing. We ask crowd workers to rewrite the toxic sentences in a non-toxic way and employ other workers to verify the rewritten versions. We ask several workers to paraphrase each of the toxic messages and yield multiple paraphrases.

Our main contribution is a new crowdsourcing setup for collecting parallel data for style transfer tasks (in our case – for detoxification task). We also release a collected parallel dataset.¹

2. Related Work

Here we discuss the approaches to creating training data for style transfer and, in particular, for detoxification task.

2.1. Style Transfer Datasets

The vast majority of datasets for style transfer are non-parallel. Thus, they consist of two or more subcorpora of a particular style. The subcorpora are not related to one another. Collecting such datasets is usually relatively easy because we do not need to label the texts for style explicitly. The style labels often already exist in the data (as in case of positive and negative reviews [12]²) or their source serves as a label (e.g. texts from Twitter can be labelled with “Twitter” style, scientific articles get a label “academic”, etc.). In these cases, data collection boils down to fetching the texts from the original sources, and the corpus size depends solely on the number of texts in these sources. A more complicated scenario is required when “style” is not a property of the data source. In the case of corpora labelled for toxicity, each text has to be manually labelled for the presence or absence of toxic content. However, even in this case, it is possible to collect a large number of texts – the size of the first Jigsaw dataset [19] contains 140,000 sentences, and the aggregated size of three Jigsaw datasets [19, 20, 21] is around 2 million unique entities.

In contrast to that, parallel corpora are more difficult to generate. For some other tasks, the parallel data may appear naturally, so that researchers do not need to perform “translation” for the research purposes. For example, translations between languages are generated simply because there is a need to transfer the information to people who do not speak the source language. Similarly, the summarization task can use abstracts of scientific papers [22], “TL;DR”

¹https://github.com/skoltech-nlp/parallel_detoxification_dataset

²<https://github.com/lijunen/Sentiment-and-Style-Transfer>

sections of social media posts [23] and summaries of news articles [24], which are created for convenience and not solely for textual data collection. There exist examples of naturally created parallel style transfer datasets. One of them is the Bible dataset [17] which exists in multiple translations of different epochs. Another example is the simple-to-regular Wikipedia dataset [25] which was generated automatically by aligning the text of Wikipedia with that of Simple Wikipedia. Likewise, a biased-to-neutral Wikipedia corpus [26] was created automatically using the information on article edits.

Besides these special cases, there exists a large style transfer dataset that was created from scratch. This is GYAFC dataset [14] which contains informal sentences and their formal versions written by crowd workers and reviewed by experts. They used Amazon Mechanical Turk to collect the data. The task was only to generate the paraphrasing and the quality control was done manually. We suggest a new pipeline that does not require manual selection of samples for the dataset and provide automatic control of the quality.

2.2. Toxicity Datasets

There exist a large number of datasets labelled for toxicity. They contain sentences and paragraphs from various social media such as Twitter [6, 27, 7], Reddit [28, 29], Facebook [30]. There exist a number of datasets based on Wikipedia talk pages – namely, Jigsaw datasets [19, 20, 21] and some others [8]. The majority of data is labelled via crowdsourcing, except for several expert-labelled corpora [6]. The labels usually include several varieties of toxicity, for example, they differentiate between offence and hate speech [27, 7]. Some datasets are restricted to a particular type of toxicity, e.g. sexism and racism [6, 7] or explore a particular type of toxicity, for example, toxic unanswerable questions [29]. The size of the datasets ranges from 9,000 [7] to 100,000 sentences [8], of which toxic sentences usually constitute 25–30%.

None of these datasets is parallel, because they were created for toxicity classification, and not for rewriting of toxic sentences. However, in works on text detoxification, the authors use these datasets either directly for training their models or to train classifiers which are then used for automatically labelling more data with (toxic or neutral) style. Nogueira dos Santos et al. [9] use the Reddit and Twitter corpora automatically labelled for toxicity using a pre-trained classifier. Tran et al. [10] collect a subset of controversial Reddit comments and label them automatically based on the presence or absence of offensive words from a manually constructed list. In the work [11] the authors use the only large dataset with manual toxicity labelling, namely, the Jigsaw dataset used in the competition of fighting the Unintended bias in toxicity detection [20]. Similarly to this work, we employ one of the Jigsaw datasets. However, our work is the first to create a parallel detoxification corpus.

3. Paraphrase Generation Pipeline

We ask crowd workers to rewrite toxic sentences so that they keep the original content and do not sound offensive. We hire workers via Yandex.Toloka³ platform. Since the phrases produced by crowd workers can be of low quality, they should not be used without revision. In [14] this

³<https://toloka.yandex.com>

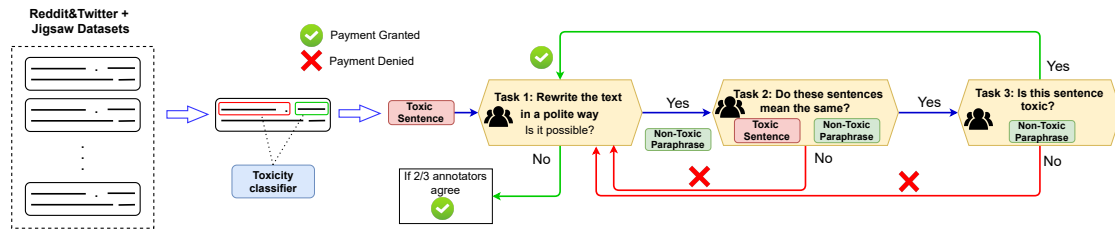


Figure 1: The pipeline of crowdsourcing for generation of detoxifying paraphrases.

revision was conducted by experts. We suggest an alternative setup where sentences are also verified by (other) crowd workers.

The objective of detoxification is to rephrase a toxic sentence so that (i) it keeps the original content and (ii) stops being toxic. Thus, we create three crowdsourcing projects: one for collecting paraphrases and another two for checking if these paraphrases conform to the two objectives. We apply the three tasks to the data sequentially. The overall pipeline is shown in Figure 1.

3.1. Task 1: Generation

In the first task, we ask users to rephrase a given sentence so that it does not sound offensive. Workers are shown a phrase and should rewrite them in a textbox (see Figure 2a).

However, not all sentences can be detoxified. First, there exist sentences that do not contain any non-toxic content — so removing toxicity from them would leave nothing from the original sentences. Consider the following examples:

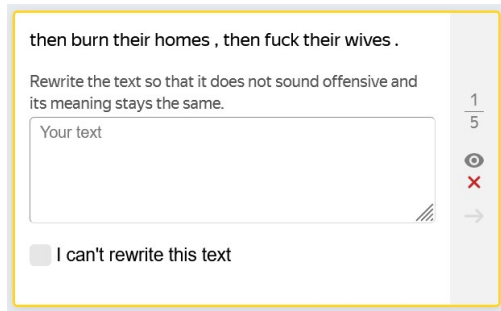
- *It sucks that you're an awful person,*
- *Maybe they should deport you back to your country, or your grandparents country.*

Their only content is an offence towards the reader, so if we detoxify them, their sense will differ substantially.

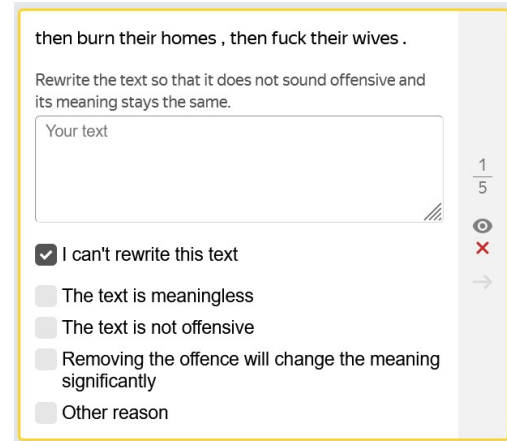
Secondly, since we pre-select sentences for detoxification semi-automatically (see Section 4.1), some of the input sentences can be non-toxic and need no detoxification. Finally, some of the given sentences can be unclear to workers due to lack of context or simply because they are meaningless.

If we do not give the worker the possibility to skip a sentence without providing a paraphrase, we can yield a large number of bad-quality paraphrases. Therefore, we extend the task interface with the option “I can’t rewrite the text”. When a user chooses it, they are shown the possible reasons for the inability to rewrite (see Figure 2b).

On the other hand, some workers can use the “I can’t rewrite” option to cheat. Since choosing this option is faster than rewriting a sentence, they can choose it too often. To avoid that, we give each input sentence to three workers and pay them for choosing the “I can’t rewrite” option only if two or more of them agree on it. Otherwise, if a worker submits a paraphrase, we check it manually in Tasks 2 and 3 and pay only for the paraphrases approved during both tasks.



(a) Generation of paraphrases.



(b) Reasons of inability to rewrite.

Figure 2: Interface of Task 1 (generation of paraphrases).

3.2. Task 2: Content Preservation Check

After having generated the paraphrases, we show them to users along with the original messages and ask if the meanings of the two sentences are close. This procedure checks if the content was preserved in the detoxified version of the input sentence and also implicitly filters out senseless outputs because they obviously do not keep the original content. The task interface is shown in Figure 3.

3.3. Task 3: Toxicity Check

The second series of paraphrases review checks if the workers succeeded in eliminating toxicity. We show users the paraphrases and ask if they contain any offence or swear words (see Figure 4).

Besides filtering out unsuitable paraphrases, Tasks 2 and 3 serve for paying for work done in Task 1. Namely, we only pay for paraphrases which the checks in Tasks 2 and 3.

The overall data collection pipeline is the following:

- We select sentences for rewriting,
- We feed the sentences to **Task 1**,
- We feed the paraphrases generated in Task 1 to **Task 2**,
- We feed the paraphrases which passed Task 2 to **Task 3**,
- We pay for paraphrases from Task 1, if they passed checks in Task 2 and Task 3,
- We pay for “I can’t rewrite” answers if two or more workers agreed on them.

4. Crowdsourcing Settings

Here we describe the settings of our crowdsourcing projects in more detail.

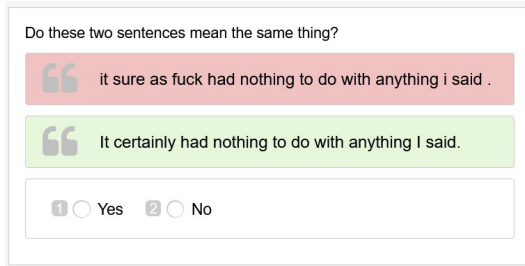


Figure 3: Interface of Task 2 (evaluation of content match).

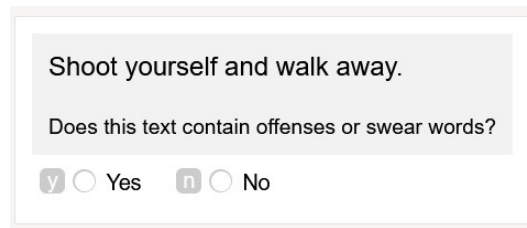


Figure 4: Interface of Task 3 (evaluation of toxicity).

4.1. Data Preprocessing

We fetch data for manual rewriting from two sources: (i) Jigsaw dataset of toxic sentences [19], and (ii) Reddit&Twitter dataset used by [9].

The Reddit&Twitter dataset was pre-processed by the authors [9]. This preprocessing included dividing original comments into sentences and automatically classifying them as toxic or non-toxic with a pre-trained classifier. The reason for this is that the original comments can be long, and generation models perform worse on longer texts. We use this Reddit&Twitter data and perform the analogous preprocessing for the Jigsaw dataset: we divide it into sentences and classify them with a pre-trained toxicity classifier.

To train a classifier, we merge the English parts of the three datasets by Jigsaw [19, 20, 21], containing around 2 million examples. We take half of this merged dataset and fine-tune a RoBERTa model [31] on it. We use the `roberta-large` model from the original repository. The classifier reaching the AUC-ROC of 0.98 and F_1 -score of 0.76 on the test set of the first Jigsaw dataset.

We select only sentences classified as toxic, whose length is between 5 and 20 words (for both datasets). The lower length limit serves for filtering out overly short sentences which are difficult to understand without context. The upper limit is for the convenience of workers who struggle with the rewriting of long sentences.

4.2. Multiple Labellings

We ask several workers to label each example. For Tasks 2 and 3, we compute the final label using the Dawid-Skene aggregation method [32] which defines the true label iteratively giving more weight to the answers of workers whose opinion more often agrees with that of other workers.

In Task 1 the number of labellings per task is 3. In Tasks 2 and 3, the number of labellings per task is defined dynamically depending on the agreement on the task. First, the minimum number of users (3 in our setting) label an example, and if their agreement is above a threshold (0.8 in our setting), the task is not sent to extra labelling. Otherwise, the system fetches up to the maximum number of labellings (5 in our setting) until the agreement reaches the threshold.

4.3. Instructions

The instruction for crowd workers says that they should rewrite the given sentences so that they **mean the same** as the original ones, but **do not sound offensive** anymore. We do not explicitly define the notion of offensiveness, because we suggest that it should be an intuitively clear concept. Instead of giving a definition of toxicity, we give examples of toxic and non-toxic sentences.

4.4. Training of Workers

For all tasks, workers have to complete a set of training tasks to be admitted to work on paid tasks. These tasks have a pre-defined correct answer. If a user answer does not match the correct one, the user is shown a hint which explains their mistake. Workers are admitted to the paid task if they completed the training with a score above a threshold: 40% correct answers for Task 1, 90% for Tasks 2 and 3. Such low acceptance score for Task 1 is due to the task complexity – there is no single answer whether the sentence can be rewritten into a civil manner or not because different people from different cultural background can have different perspective on the toxicity. At the same time, even such a low score allows us to ban the user that just select only one answer, for example, and do not put effort into thorough task completeness. Task 2 and 3 have more exact answer and, as a result, have a higher threshold of acceptance.

In addition to that, we provide extra training tasks during labelling. Training tasks are added to labelling examples. If a worker makes an error in such tasks, they are shown an explanation. Workers are not penalised for making errors. We need these tasks to reinforce the worker's understanding of the task.

4.5. Quality Control

To control the quality, we occasionally add control tasks during labelling. These tasks have a pre-defined correct answer. If a worker makes too many mistakes in these tasks, they are banned. In our projects, a worker is banned if they make 3 or more mistakes in the latest 4 control tasks.

Another means of control that we use is the agreement between users. If a user disagrees with the majority (3 or more) of users in over 66% cases, this user is banned.

4.6. Payment

In Yandex.Toloka, a task page shown to workers can contain multiple tasks. Therefore, customers set the payment per task page. In Task 1, a page contains 4 tasks for labelling and one control task and costs \$0.02. In Tasks 2 and 3, a page consists of 10 labelling tasks, 1 to 2 training tasks and 1 to 2 control tasks and costs \$0.01. The number of training and control tasks is defined by the agreement of the worker with others.

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

Original	Paraphrase
Reddit	
this is scaring the shit out of me.	This is really scaring me.
this is a joke , are you all fucking retards?	This is a joke, are you all crazy?
everybody is such a fucking pussy.	Everybody is acting cowardly.
did you think i was going to sit back and let some imbecile spew hatred towards me?	Did you think I was going to let an unreasonable person hate me?
calm the fuck down, cnn.	Please calm your nerves, CNN.
why is the scale of that graph so fucked up?	Why is the scale of that graph not placed rightly.
Jigsaw	
the chinese came out of this looking like lying, petulant children.	The chinese came out of this looking like dishonest, bad tempered children.
youre a libby so dont talk ridiculous like you would vote any other way.	You're a liberal so you wouldn't vote other way.
she is a lying disgrace who refuses to accept that she lost.	She is lying because she does not want to accept she has lost.
all you trump clowns are seriously messed up.	Trumps voters are seriously mislead.
if you think so, you are making a clown out of yourself.	It is wrong of you to think that way.

Table 1
Examples of detoxifying paraphrases from our corpus.

Figure 1: the green arrows indicate variants where we pay for Task 1 and the red arrows show when it is not paid.

5. Data and Pipeline Analysis

We tested a pipeline for collecting parallel detoxification data. As a proof of concept, we collected a dataset with paraphrases for 1,244 toxic sentences, each having 1-5 paraphrase variants, resulting in 5,228 paraphrases overall. The examples of generated paraphrases are given in Table 1.

5.1. Data Statistics

The statistics of the paraphrases written by crowd workers are presented in Table 2. For both Jigsaw and Reddit&Twitter datasets, we used the equal number of sentences (1,200) as the input to the crowdsourcing pipeline from Figure 1. From this input data, 675 samples from Jigsaw (56%) received at least one good paraphrase, while for the Reddit dataset this amount is smaller — 569 samples (47%).

Source Dataset	# Input Samples	# Unique Samples Covered	# Paraphrases Total	Task2 Confidence $\geq 90\%$	Task3 Confidence $\geq 90\%$	Cost per 1,000 inputs	Cost per final sample
Jigsaw	1,200	675	2,141	92%	92%	\$60	\$0.10
Reddit	1,200	569	3,087	95%	92%	\$60	\$0.10
Total	2,400	1,244	5,228	93%	92%	\$60	\$0.10

Table 2
Statistics of the crowdsourcing experiments and final datasets.

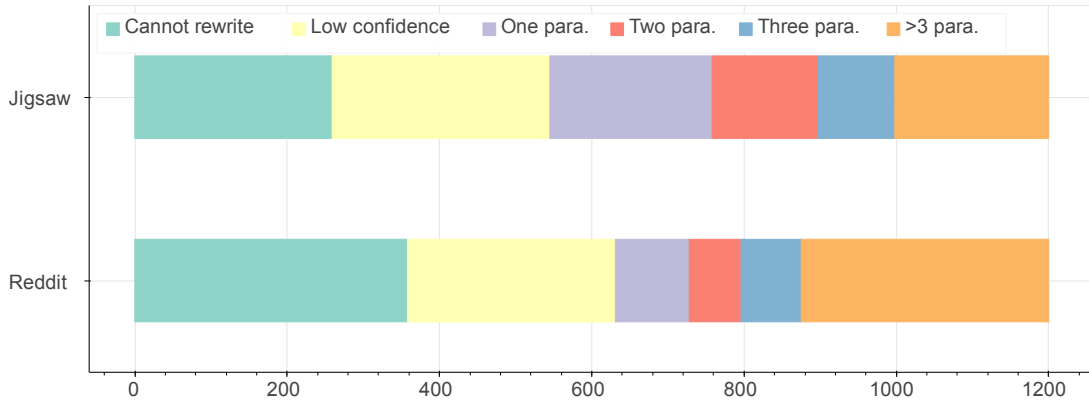


Figure 5: Distribution of results of detoxification experiments.

Some of the input samples could not be paraphrased or all paraphrases generated for them were rejected during verification. Others yielded one or more paraphrases. The distribution of the different results is shown in Figure 5. The pictures for Jigsaw and Reddit datasets differ. Besides the mentioned difference in the number of input samples which received paraphrases, Reddit has a larger number of sentences with more than one paraphrase. These sentences are probably easier to detoxify because multiple workers succeeded in providing paraphrases for them. Reddit has more of such sentences. The annotators consistency is quite high (Table 2, columns 6 and 7). For Task2 and Task3 for both datasets $\geq 92\%$ of all marked up samples have the confidence in the decision of $\geq 90\%$. We compute the cost of manual annotation. The cost per input sample is the cost of generating 1-5 paraphrases for the sample and checking these paraphrases for toxicity and content matching. The cost of the final sentence pair includes the cost of processing the inputs for which no good paraphrases were generated. Thus, this parameter depends not only on crowdsourcing settings but also on the dataset, because the percentage of detoxifiable sentences may vary for different data sources.

Multiple paraphrases can give us a better insight into the phenomenon of toxicity. Namely, by comparing multiple paraphrases of one sentence we can detect its toxicity regions with higher confidence. While a single paraphrase can be uninformative (e.g. it can fully rephrase a sentence where toxicity could be eliminated by replacing several words), multiple paraphrases show a more objective picture. Table 3 shows examples of such cases. The colour is brighter for words

Original	to think that she's been doing that kind of shit for yeaaaars
Paraphrase	<p>her behaviour has been constant for years.</p> <p>to think that she's been doing that kind of thing for years.</p> <p>to think that she's been doing that for years.</p> <p>to think she's been doing that for years.</p> <p>to think that she's been doing this kind of thing for years</p> <p>she is been doing those things for years</p> <p>to think that she's been doing it for years.</p>
Original	this is why the country is fucked up.
Paraphrase	<p>this is why the country does not going well.</p> <p>this is why the country is not growing</p> <p>country is suffering because of this</p> <p>this is why the country is bad.</p> <p>this is why this country is messed up.</p> <p>that's why the country is lost</p> <p>that's why this country is in such a mess.</p> <p>this country is bad</p> <p>this may be the reason why our country is not so progressive</p> <p>this is why the country isn't developing.</p> <p>this is why the county has problems.</p>

Table 3

Determination of the most toxic regions of a sentence based on multiple detoxifications.

that are absent in the detoxified versions more often. It can be seen that while some non-toxic words can occasionally be removed, the joint statistics of multiple paraphrases identifies the regions of toxicity. This token-level information can be useful for toxicity classification as well. As shown in [33], the rationales (highlighted phrases that justify the class label assigned to a sentence) improve the performance of text classifiers and their explainability. The token-level degree of toxicity can serve as a rationale for sentence-level classifiers.

Also, it can be seen from Tables 1 and 3 that there can be some grammatical mistakes in both original and generated sentences. The reason for that is that the majority of English content in the Internet is produced by people for whom English is a second language. The same situation about language we have for the annotators in our task. So, we consider these mistakes as natural for both original and detoxified sentences.

5.2. Crowdsourcers Performance

Throughout data collection, up to 2,000 crowd workers participated in our experiments (in each task). However, around 1/2 (Tasks 2 and 3) to 3/4 (Task 1) of them were banned for cheating (too fast answers, failed captcha checks, errors in control questions).

We can measure the performance of crowd workers with a number of parameters. These are their agreement computed as the average confidence computed with Dawid-Skene aggregation method [32] and their performance on pre-labelling training and control tasks. In the most

challenging Task 1 the performance on the control questions was 32.8% of correct answers. In Tasks 2 and 3 these numbers were 63.4% and 86.9%, respectively. Thus, Task 3 was the easiest.

This is corroborated by the fact that users completed this task faster. It took them on average 1 minute 31 seconds to label 13-15 sentences (one task page) as toxic or safe. For Task 2 this number is 3 minutes 8 seconds (for 13-15 tasks), and one Task 1 page was completed on average in 5 minutes 17 seconds (5 tasks).

In addition to that, Task 3 was rated higher by crowd workers than the other two tasks. In Yandex.Toloka workers can rate projects by four parameters:

- how interesting the task is (formulated as “Would you complete this task in the future?”),
- clarity of instructions,
- task interface,
- communication with the requester.

Task 3 got an average score of 4.77, whereas Tasks 1 and 2 got scores of 4.51 and 4.64, respectively.

6. Conclusions and Future Work

We describe a crowdsourcing setup for the collection of parallel data for the detoxification task. The data consists of pairs of sentences, one of them is toxic, and the other one has the same meaning but is not offensive. In our setup workers write paraphrases for toxic sentences and verify the content preservation and the absence of toxicity.

We collected 1-5 manually written paraphrases for over 1,200 toxic sentences from Reddit, Twitter, and Wikipedia discussion pages. This corpus is relatively small. However, we believe that even this dataset can be used for supervised training of a detoxification model, for example, for fine-tuning of a large language model, such as GPT-2 [34] or T5 [35]. Besides, it can be used for the training of a token-level toxicity classifier and as a source of extra information in a sentence-level toxicity classifier.

The most evident direction of future work is the collection of a larger dataset and investigation of the optimal number of parallel examples for the training of a well-performing detoxification model. In addition to that, we would like to further improve our crowdsourcing pipeline by finding the optimal number of labellings per sample and compare the cost and efficiency of setups with one and multiple paraphrases.

We would also like to further research the usefulness of multiple paraphrases for a single toxic message for the training and evaluation of detoxification models. Finally, our experiments show that data from different sources can have different properties important for parallel data collection. Namely, the Reddit dataset has more sentences that can be paraphrased. It would be useful for further data collection to investigate the properties of different datasets.

References

- [1] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on

- Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447. URL: <https://www.aclweb.org/anthology/2020.semeval-1.188>.
- [2] A. G. D’Sa, I. Illina, D. Fohr, Towards non-toxic landscapes: Automatic toxic comment detection using DNN, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 21–25. URL: <https://www.aclweb.org/anthology/2020.trac-1.4>.
- [3] X. Han, Y. Tsvetkov, Fortifying toxic speech detectors against veiled toxicity, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7732–7739. URL: <https://www.aclweb.org/anthology/2020.emnlp-main.622>. doi:10.18653/v1/2020.emnlp-main.622.
- [4] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: <https://aclanthology.org/W17-1101>. doi:10.18653/v1/W17-1101.
- [5] A. Pelicon, R. Shekhar, M. Martinc, B. Škrlj, M. Purver, S. Pollak, Zero-shot cross-lingual content filtering: Offensive language and hate speech detection, in: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, Association for Computational Linguistics, Online, 2021, pp. 30–34. URL: <https://aclanthology.org/2021.hackashop-1.5>.
- [6] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: <https://aclanthology.org/N16-2013>. doi:10.18653/v1/N16-2013.
- [7] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: <https://aclanthology.org/S19-2007>. doi:10.18653/v1/S19-2007.
- [8] E. Brassard-Gourdeau, R. Khoury, Subversive toxicity detection using sentiment information, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1–10. URL: <https://aclanthology.org/W19-3501>. doi:10.18653/v1/W19-3501.
- [9] C. Nogueira dos Santos, I. Melnyk, I. Padhi, 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 189–194. URL: <https://aclanthology.org/P18-2031>. doi:10.18653/v1/P18-2031.
- [10] M. Tran, Y. Zhang, M. Soleymani, Towards a friendly online community: An unsupervised style transfer framework for profanity redaction, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 2107–2114. URL: <https://aclanthology.org/2020.coling-main.190>. doi:10.18653/v1/2020.coling-main.190.
- [11] L. Laugier, J. Pavlopoulos, J. Sorensen, L. Dixon, 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, Association for Computational Linguistics, Online, 2021, pp. 1442–1461. URL: <https://aclanthology.org/2021.eacl-main.124>.
- [12] J. Li, R. Jia, H. He, P. Liang, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1865–1874. URL: <https://aclanthology.org/N18-1169>. doi:10.18653/v1/N18-1169.
- [13] K. Chawla, B. V. Srinivasan, N. Chhaya, Generating formality-tuned summaries using input-dependent rewards, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 833–842. URL: <https://aclanthology.org/K19-1078>. doi:10.18653/v1/K19-1078.
- [14] S. Rao, J. Tetreault, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 129–140. URL: <https://aclanthology.org/N18-1012>. doi:10.18653/v1/N18-1012.
- [15] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts, A computational approach to politeness with application to social factors, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 250–259. URL: <https://aclanthology.org/P13-1025>.
- [16] A. Madaan, A. Setlur, T. Parekh, B. Poczós, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, S. Prabhunoye, Politeness transfer: A tag and generate approach, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1869–1881. URL: <https://aclanthology.org/2020.acl-main.169>. doi:10.18653/v1/2020.acl-main.169.
- [17] K. Carlson, A. Riddell, D. Rockmore, Evaluating prose style transfer with the bible, *Royal Society Open Science* 5 (2018).
- [18] D. Dementieva, D. Moskovskiy, V. Logacheva, D. O. Kozlova, N. Semenov, A. Panchenko, Methods for detoxification of texts for the Russian language, in: Proceedings of the International Conference “Dialogue 2021”, Moscow, Russia, 2021. doi:99.9999/woot07-S422.
- [19] Jigsaw, Toxic comment classification challenge, <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>, 2018. Accessed: 2021-03-01.
- [20] Jigsaw, Jigsaw unintended bias in toxicity classification, <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>, 2019. Accessed: 2021-03-01.
- [21] Jigsaw, Jigsaw multilingual toxic comment classification, <https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification>, 2020. Accessed: 2021-03-01.
- [22] I. Cachola, K. Lo, A. Cohan, D. Weld, TLDR: Extreme summarization of scientific documents, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4766–4777. URL: <https://aclanthology.org/>

- 2020.findings-emnlp.428. doi:10.18653/v1/2020.findings-emnlp.428.
- [23] M. Völske, M. Potthast, S. Syed, B. Stein, TL;DR: Mining Reddit to learn automatic summarization, in: Proceedings of the Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 59–63. URL: <https://aclanthology.org/W17-4508>. doi:10.18653/v1/W17-4508.
- [24] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 28, Curran Associates, Inc., 2015. URL: <https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf>.
- [25] Z. Zhu, D. Bernhard, I. Gurevych, A monolingual tree-based translation model for sentence simplification, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 1353–1361. URL: <https://aclanthology.org/C10-1152>.
- [26] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, D. Yang, Automatically neutralizing subjective bias in text, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 480–489. URL: <https://aaai.org/ojs/index.php/AAAI/article/view/5385>.
- [27] T. Davidson, D. Warmsley, M. W. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, AAAI Press, 2017, pp. 512–515. URL: <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665>.
- [28] J. Kurrek, H. M. Saleem, D. Ruths, Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 138–149. URL: <https://aclanthology.org/2020.alw-1.17>. doi:10.18653/v1/2020.alw-1.17.
- [29] S. Bagga, A. Piper, D. Ruths, “are you kidding me?”: Detecting unpalatable questions on Reddit, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2083–2099. URL: <https://aclanthology.org/2021.eacl-main.179>.
- [30] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1–11. URL: <https://aclanthology.org/W18-4401>.
- [31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [32] A. P. Dawid, A. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, Journal of The Royal Statistical Society Series C-applied Statistics 28 (1979) 20–28.
- [33] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A

benchmark dataset for explainable hate speech detection, 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, AAAI Press, 2021, pp. 14867–14875. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/17745>.

- [34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
- [35] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.