www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Detoxifying Text with the Use of Massive Pre-trained Neural Networks

[1] Sankham Surendra, [2] J.A.Paulson,

[1,] PG Student, Department of CSE, Varaprasad Reddy Institute of Technology, Sattenapalli, Kantepudi (Village)

[2] Associate Professor, Department of CSE, Varaprasad Reddy Institute of Technology, Sattenapalli, Kantepudi (Village)

Abstract

Two new unsupervised algorithms for removing harmful language from text are introduced. First, we employ paraphrase models to execute style transfer. Second, we guide the creation process using modest style conditional language models. These two current concepts come together in our first technique. In order to preserve the text's meaning while removing poisonous elements, we use a high-performing paraphraser that is guided by language models educated on style. Our second approach makes use of BERT to swap out harmful terms with their less offending counterparts. We provide flexibility to the technique by letting BERT replace mask tokens with a configurable amount of words, making it more versatile. At last, we provide the first comprehensive comparison of style transfer methods for the elimination of toxicity. Several different approaches to style transmission are compared using our models. utilizing a mix of unsupervised style transfer measures, the models are tested without utilizing any references. Our two proposed methods both provide novel SOTA outcomes.

1 Introduction

Zampieri et al. (2020), D'Sa et al. (2020), and Han and Tsvetkov (2020) are all part of the ongoing study on toxicity detection in user writings. Although the job of automatically rewriting objectionable information did not get much attention, it might have many practical uses, such as helping to improve the online environment by suggesting a more neutral version of an emotive statement. Dos Santos et al. (2018), Tran et al. (2020), and Laugier et al. (2021) are among the studies on text detoxification that frame this goal as style transfer. Rewriting text while changing one or more properties that make up the "style"—for example, authorship, emotion, or politeness level—is often understood as the style transfer job (Voigt et al., 2018; Shen et al., 2017; Madaan et al., 2020). Although the intention is to keep the text intact, altering the style characteristics may often drastically alter the meaning of a sentence. 1 In reality, the objective of several style transfer models is to change the style of a phrase such that it is substantially similar to another statement on the same subject but with a different style. 2 We argue that detoxification, unlike many other style transfer jobs like sentiment transfer, requires a distinct approach in order to better preserve the original meaning. Two text detoxification methods with further control for content retention are shown here. It is possible to completely regenerate the input using the first model, ParaGeDi. It relies on two concepts: first, that a class-conditioned LM may externally regulate the output of a generation model (Krause et al., 2020); and second, that paraphrasing can be used to mimic a style transfer job (Kr ishna et al., 2020). Using a paraphraser model as its foundation, ParaGeDi makes it a point to maintain the original sentence's meaning. The second method, CondBERT, is based on the pointwise editing setup and was motivated by Wu et al. (2019a). It finds harmful spans in the phrase and utilizes BERT to replace them with non-toxic ones. on maintain the semantic closeness, we feed BERT the original text and adjust its hypothesis ranking according on how similar the replacement words are to the original. It is worth noting that BERT may effectively transition from a toxic to a typical text style without requiring class-conditional pre-training. Additionally, we compare our novel models to baselines and state-of-the-art methods in a large-scale assessment of style transfer models on detoxifying task. Our data and code are made public. 3 Here are the things that we have contributed: Our two new detoxifying methods, ParaGeDi (which stands for "paraphrasing GeDi") and Cond BERT (conditional BERT), are based on pre-trained neural language models. • We provide the detoxification dataset and evaluate these models by comparing them to

other state-of-the-art algorithms for text detoxification and sentiment transfer. • By mining the ParaNMT dataset for pairings of hazardous and safe sentences, we construct an English parallel corpus tailored to the detoxification task (Wieting and Gimpel, 2018). It may enhance our top-performing models even more, as we demonstrate.

2 Related Work

Using a trained encoder decoder model to "translate" a source phrase into the target style is one of the most easy techniques to solve a style transfer challenge (Rao and Tetreault, 2018). Given that both the source and destination languages are identical, pre-trained LMs like GPT-2 may be used for this purpose. By fine-tuning them on relatively small parallel corpora, they provide satisfactory results (Wang et al., 2019). (Radford et al., 2019). Unfortunately, there is a dearth of suitably huge parallel data, hence this strategy is hardly used. All of the other models that are mentioned here were trained without any supervision. A Model for Pointwise Editing Keeping the sentence as-is and changing only the words linked to the style is a simple yet effective way to shift styles. The first effort to carry out such a transfer was the Delete-Retrieve-Generate (DRG) framework (Li et al., 2018). It suggests four approaches predicated on this idea. Among these, two do quite well with our data: Finding DRGs Only obtains a sentence with the opposite style which is similar to the original sentence and returns it, and DRG TemplateBased takes the style attributes from it and inserts them into the original phrase. Methods for identifying style markers and retrieving re placements are crucial to the performance in this case. Although frequency analysis was used to identify style-related words in the original article, attention weights have also been used in other publications (Sudhakar et al., 2019). Another option is to employ Masked Language Modelling (MLM) for style transfer. An MLM that has been trained on a dataset that contains style labels will choose a new term to use according on both the context and the style label. Mask &Infill is one such model (Wu et al., 2019b). Our proposed CondBERT technique is the most comparable to it. Nevertheless, Cond BERT accomplishes more style and content preservation control and can substitute multiple words. Malmi et al. (2020) outlines an additional model that is comparable to this one. In this more convoluted setup, two MLMs trained on separate style corpora for each form replacement work together. Comprehensive Frameworks However, end-to-end structures for style transfer do exist, which is different from these approaches. After encoding the original language, they insert a new style into the concealed representation by manipulating it, and finally, they decode it. If you look closely enough, you can see that some of them separate the concealed representation from the content and style representations (John et al., 2019). According to Hu et al. (2017), the other ones make the encoder convey material that is not style dependent. Another option is the DualRL model developed by Luo et al. (2019), which directly transfers the style from the source to the target. This job may be executed in tandem with the dual task, which enables models to train even in the absence of parallel data. He et al.'s (2020) Deep Latent Sequence Model (DLSM) model trains models for both the primary and secondary tasks simultaneously using amortized variational inference. According to Lee (2020), the Stable Style Transformer (SST) technique incorporates the cross-entropy of a pretrained style classifier as an extra discriminative loss to train a pair of sequence-to-sequence transformers for primal and dual tasks. Kr ishna et al.'s (2020) Style Transfer as Paraphrase (STRAP) approach considers a style transfer model to be a paraphraser that imbues a piece of text with characteristics of a certain style. By using a pre-trained general-purpose paraphraser to convert style-marked texts to neutral, the authors generate pseudo-parallel datasets. From these, they train sequence-to-sequence models. These approaches are conceptually comparable to our ParaGeDi model. But this approach differs in that the style is imposed on the generator by another model, rather than being fused into the model or a sentence representation. Clearing the System The process of text detoxification is a novel style transfer assignment. An end-to-end seq2seq model trained on a non-parallel cortex with autoencoder loss, style categorization loss, and cycle-consistency loss was the first study on this issue by (dos Santos et al., 2018). A more recent study by Tran et al. (2020) employs a series of models in a pipeline: first, a search engine locates non-toxic phrases that are comparable to the provided toxic ones; second, an MLM completes the missing parts of the discovered sentences; and last, a seq2seq model improves the produced sentence's fluency. Finally, sentences are detoxified by Laugier et al. (2021) using T5 as a denoising autoencoder with extra cycle-consistency loss. In their respective studies, Dathathri et al. (2020) and Krause et al. (2020) tackle the same issue: how to stop a language model from producing harmful content. They are under no obligation to maintain the original intent of the supplied content. Nevertheless, our findings demonstrate that the concept of using a discriminator to regulate an LMduringgeneration may be used to style transmission.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

3 Paraphrasing GeDi Model

A language model trained on certain text attributes, such as style or theme, directs the newly suggested GeDi model (Krause et al., 2020) to generate text from beginning. We improve upon this paradigm by adding the ability to paraphrase the user-supplied content.
GeDi 3.1
Both the generation model (GPT-2) and the discrimination model—also a GPT-2—were originally part of the GeDi model. The discrimination model was trained using sentences that included extra sentence-level style labeling, meaning that the style label was appended to each phrase during training. The discriminating model is then trained to learn word distributions that are conditional on a certain label. The primary model PLM uses an extra class-conditional language model PD and the Bayes rule to modify the distribution of the next token anticipated at each generation step:

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

Attribute c, which might be anything from toxicity to emotion, is one of the C classes; xt is the present token, x\t is the text prefix, and c is the desired attribute. A combination of the Bayes rule and the supplementary class-conditional language model (PCC) is used to compute the second term, while the primary language model (PLM) produces the first term. Tokens that are more prevalent in the selected style of writing are therefore given a larger probability:

3.2 ParaGeDi

In order to enable GeDi to preserve the meaning of the input text, we replace the regular language

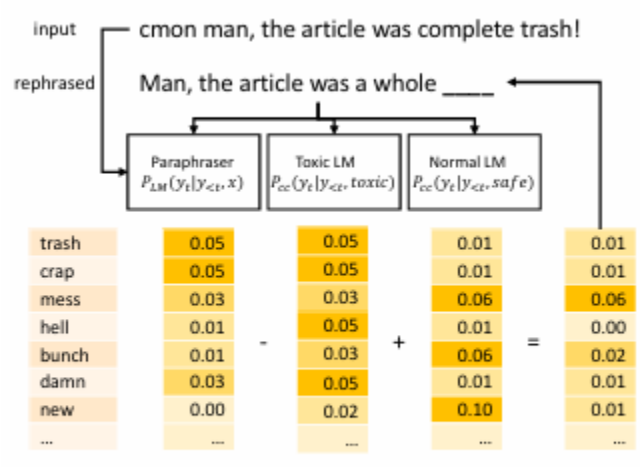

Figure 1: The overview of ParaGeDi model.

simulate it using a model that can paraphrase. The following likelihood may be modelled by ParaGeDi given the following: x is the original text, y is the produced text of length T, and c is the desired style.

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

Since x and y should be conditions on the class probability, the final step is an approximation. Our paraphraser model needs a parallel corpus for training, but we can separate it from the style model, which simply needs texts

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

with style labels—not necessarily parallel—by using this approximation, even if it isn't entirely warranted. It is possible to train the paraphraser and style model separately. In addition, any paraphraser that has the same vocabulary as the class-conditional LM may be used. A third, optional, part of this model is a reranker, which is an external model that gives more weight to the style-related hypotheses produced by the style LM-guided paraphraser. Our reranker uses a pre-trained toxicity classifier to choose the ParaGeDi model's least harmful hypothesis. The process flow of our model is shown in Figure 1. The following is the training for ParaGeDi. The destruction of it The generating loss LG is used in LM training, and the discriminative loss LD further isolates classes from one another; these two losses are linearly combined to form LParaGeDi.

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

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

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

To make the model better, we include many inferencing strategies that boost the accuracy of style transfer and content retention. A heuristic from the first GeDi model is used first. During generation, we bias the discriminator towards the right class by raising the conditional LM probability to the power of w > 1.

4 Conditional BERT Model

Since BERT has been taught to fill in gaps ("masked LM"; Devlin et al., 2019), we may utilize it to replace harmful words with non-toxic ones. For the purpose of data augmentation, Wu et al. (2019a) proposed this strategy. By locating source-style terms and replacing them with the [MASK] token, the authors instruct the BERT model to insert new words of the target style into the corresponding spaces. By substituting trainable style embeddings for the segmentation embeddings in the original BERT, the authors fine-tune BERT on a style-labeled dataset, therefore pushing BERT towards the desired style.
We modify this model so it may be used for detoxification. We choose the toxic phrases, whereas the original conditional BERT model masked them at random. One approach is to manually compile a list of offensive and poisonous terms; another is to train a word-level toxicity classifier.
We employ a technique that doesn't call for any further information or manual intervention.

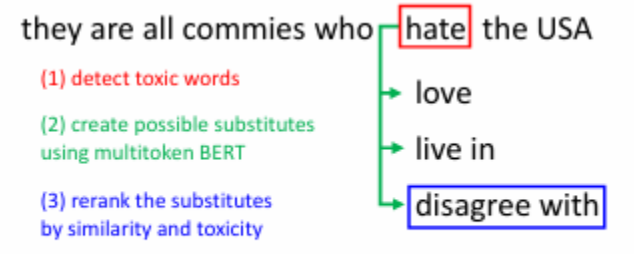

Figure 2: The overview of the CondBERT model.

Toxic words are those whose toxicity scores are greater than a certain threshold, where s1 s2 sn are the scores of all the words in the phrase and tmin = 02 is the lowest toxicity score. This threshold is determined by computing the toxicity score for each word in the sentence. By adjusting the threshold, we can stabilize the harmful word proportion in a phrase, preventing instances when either too many or no words are classified as toxic. We use the content preservation algorithms proposed by Arefyev et al. (2020) to ensure that the replacement term retains its meaning: (i) Keep the original tokens intact rather than masking them before replacing; (ii) Sort the

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

replacement words proposed by BERT according to how similar their embedding is to the original word's embedding.

Despite using class-specific phrase embeddings, conditional BERT often forecasts harmful words, perhaps prioritizing context above the target class's embeddings. We determine the toxicity of each token in the BERT vocabulary and penalize the estimated probability of tokens with positive toxicities such that the model can only construct non-toxic terms.

5. Results

You can see how well each model did in the tests in Table 1. There are alternative models that perform better than ParaGeDi and CondBERT. Heuristics directed at the metric's constituent parts account for CondBERT's performance: (i) in order to ensure a high ACC score, it is penalized for producing toxic tokens; (ii) in order to increase the overall SIM score, over 80% of tokens remain unchanged and replacements are chosen based on how similar they are to the original words; (iii) MLM is pre-trained to increase FL by replacing masked tokens with plausible alternatives. Since generation is a more effective technique for text naturalness than pointwise corrections, ParaGeDi has somewhat greater fluency but lags behind in similarity. Mask&Infill, which follows the same logic as CondBERT, is our models' main rival. However, a significant reduction in fluency and a little loss in style transfer accuracy are outcomes of some engineering choices, such as masking all words simultaneously. When compared to the two most basic (DRG) models, TemplateBased and RetrieveOnly, many more complex models actually perform worse. RetrieveOnly produces high levels of similarity and style transfer accuracy by extracting actual, non-toxic sentences from the training data, whereas TemplateBased accomplishes a high level of similarity by preserving the majority of the original sentence. complete text re-generation (rather than pointwise corrections) using the DLSM and SST models. They relied on a short dataset, which is why their fluency ratings were low. On the other hand, STRAP is more fluent as it utilizes the bigger pseudo-parallel data set, which it also uses to generate the sentence. The capacity of MT to detoxify is another discovery. On the other hand, its quality is inversely linked to its quantity: the En Ig En only detoxifies 37% of sentences, although it gets poor ratings for SIM and FL. On the other hand, En Fr En produces superior results while retaining the majority of the initial characteristics, such as toxicity. The T5 paraphraser is no different. In contrast, our research only required 200 parallel words to teach the GPT-2 model to detoxify. We propose that training it on a bigger parallel dataset may improve its performance, even if it performs worse than many other models. The best-performing models' paraphrases are shown in Table 2. You may find more examples in Appendix F and qualitative analysis in Appendix E. Choose the Parameters (5.7) A number of heuristics and parameters are used by our models. In order to determine their utility, we conduct an ablation research. It turns out that CondBERT's most important properties are toxicity penalty, which increases style strength, and multiword replacement, which guarantees good fluency. However, quality is unaffected by controlling similarity or masking all tokens simultaneously. The Cond BERT ablation investigation is further detailed in Appendix B. One training hyperparameter regulates the strength of ParaGeDi's discriminative loss. We find that its value little affects the overall quality; for example, J's value drops to zero at = 1, indicating no generative loss (refer to Figure 3). Using a word probability higher limit enhances similarity, disabling beam search lowers fluency, and adjusting the style intensity control affects style correctness. Beam size, smoothing, and reranking, in contrast, have no effect on the model's accuracy. The ParaGeDi model's ablation investigation is detailed in Appendix C.

6 MiningaParallel Detoxifying Corpus

The STRAP model (Krishna et al., 2020) presumes that a skilled paraphraser can render a text with stylistic markings neutral. However, as we have shown in our tests, a para

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

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

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

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

Table 2: Examples of detoxification by different models.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

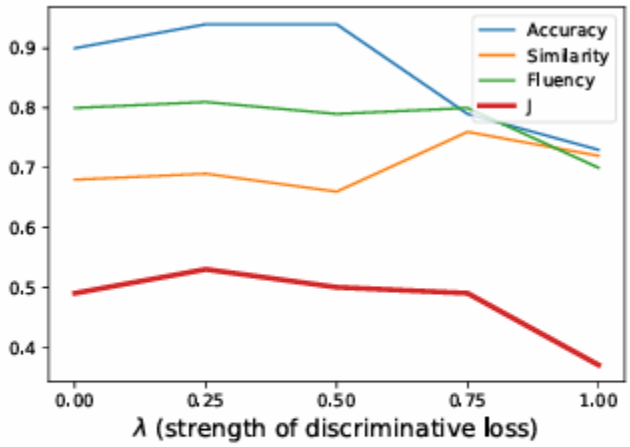

Figure 3: Performance of ParaGeDi with the varying λ parameter (greater λ corresponds to the stronger influence of the discriminative loss and smaller λ means the stronger influence of the generative loss).

We propose that it is feasible to discover infrequent detoxifying sentence pairings in a large parallel dataset of paraphrases, as phrasers and MT models are lousy detoxifiers on their own (see Section 5.6). Setting up experimental To put this theory to the test, we categorize the phrases from the ParaNMT to

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

Table 3: Comparison of paraphrasers for ParaGeDi.

500,000 paraphrase pairings where one sentence is more hazardous than the other were used with our toxicity classifier, as stated in Section 5.1 (for further information on the data collecting procedure, please refer to Appendix D). The dataset was compiled by Wieting and Gimpel (2018). Check out the normal paraphraser from Section 5.4, which is fine-tuned on a random subset of ParaNMT, and see how it stacks up against its variant, which is fine-tuned on the toxic/safe parallel paraphrase corpus. Additionally, we compare the overall performance of both paraphrasers by plugging them into the ParaGeDi model. The outcomes may be shown in Table 3.

7 HumanEvaluation of Detoxification

Automatic reference-free assessment is quick and inexpensive, but it may not be accurate. Classifiers for toxicity and fluency could make mistakes and provide inaccurate results. There was a limited correlation between human judgments and the embedding distance, a metric for content retention (Yamshchikov et al., 2021). That is why we do a manual evaluation of the top models.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Setting up experimental We aim for maximum resemblance to the automated assessment in the design of our manual evaluation setup. When assessing our models, we always use the same three criteria: ACCm for style, SIMm for content similarity, and FLm for fluency. A ternary scale: {0, 0.5, 1} corresponds to a horrible sentence, a partly acceptable sentence, and a totally acceptable sentence, and we utilize this scale for all metrics. To assess the models, we enlist the help of five annotators. Annotators are researchers in natural language processing who have a master's degree or above and who speak English well. We scheduled a preliminary round to establish mutual understanding of annotation before beginning. After three annotators have reviewed each phrase, the final score is calculated by taking the average of their ratings. Using Krippendorff's metric, we assess the level of agreement between annotators. The scores we obtained for style correctness (0.42), content preservation (0.31), and fluency (0.52), indicating a reasonable agreement for style and fluency annotation, and a poor agreement for content annotation.

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

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

by all of the models that were considered. To eliminate unnecessary or fluff, the evaluation's input (toxic) phrases were hand-picked (without taking the outputs into account). We reported the average score after triple-tagging each sample to make up for the poor agreement amongst the annotators. Results Discussion Table 4 displays the results of the models as assessed by humans. Summarized sentence scores constitute the model scores. We call the combined quality score of the three criteria Jm. Jm at the sentence level is the product of the ACCm, SIMm, and FLm scores, and the model Jm scores are the mean of these values. The results of this manual examination confirm that our models are better than the Mask&Infill model. Not only that, but it verifies that there is little to no difference between our two models. Even while ParaGeDi does better than CondBERT on every parameter, the gap between the two is only statistically significant for FLm. We also looked at how well the automated measures mirror human opinions, in addition to the assessment itself. We do this by calculating their Spearman's correlation score with human evaluations (see to Table 6 for details). We take into account the ACC and ACC-soft versions of the toxicity classifier, which yield confidence instead of a binary label, when thinking about style. In terms of content, we evaluate the original and detoxified sentences using the SIM (embedded similarity) and BLEU scores. Many research on style transfer and other generation tasks employ the perplexity of the GPT-2 language model (PPL) and the linguistic acceptability classifier (FL) to evaluate fluency (Radford et al., 2019).

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

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

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

ACC_m		SIM_m		FL_m	
ACC-soft	0.59	SIM	0.34	FL	0.54
ACC	0.51	BLEU	0.19	PPL	0.45

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

compares somewhat to human evaluations of style and fluency measurements generated automatically. Confirming the critique of perplexity as a fluency measure, it comes out that the confidence of style classifier is a superior style accuracy metric than a binary classifier. Additionally, the ac ceptability classifier performs better than perplexity.

8 Sentiment Transfer Experiments

Because text detoxification is a relatively new style transfer problem, comparing our models to others in the field is challenging. As a result, we switch gears and test out sentiment transfer as a domain. Setup for the Experiment Using the Yelp reviews dataset (Li et al., 2018), we train ParaGeDi and CondBERT and evaluate them alongside other models (see Section 2), including Mask&Infill, SST, DRG-TemplateBased, DRG-RetrieveOnly, and Du alRL. We use the results of previous models developed by their creators and fine-tune the hyperparameters of ParaGeDi and CondBERT on the Yelp development set. Just as in our detoxifying studies, we use the J to assess the models. As described in Section 5.1, we train two sentiment classifiers on two separate sections of the Yelp dataset in order to evaluate the style transfer accuracy. We use one for drawing conclusions and another for assessing quality. Using the human references supplied by Li et al. (2018), we additionally calculate the BLEU score. Table 5 displays the data, the average of which is displayed for both transfer directions. Results Discussion When comparing models, ParaGedi performs better in terms of J. The other models still can't make natural-sounding writings as they either learn to make up texts from beginning or focus on replacing certain words. The only rival that integrates pre-trained models with complete regeneration is the ParaGeDi model. Detoxification and style transfer for other domains need new approaches, since the CondBERT model performs poorly on this job. However, the BLEU score casts doubt on this outcome. In terms of performance relative to human references, DualRL, Mask&Infill, and our CondBERT are the top three MLM-based models. The referenceless metrics are also called into question in the reference hu man replies assessment. The performance of the classifier is the first constraint on the ACC score. Due to the low reliability of ACC values over 0.81 and the fact that it only provides 0.81 for ten phrases that were assumed 100% correctly written by hand, it is unreasonable to assume that minor variations in ACC are significant. Overall, ParaGeDi may still be deemed a powerful style transfer model due to the closeness of the human responses' scores to those of Mask&Infill. Metrics are unable to differentiate between the models at this level, thus more precise evaluations should be conducted by humans.

9 Conclusion

Two models for style transfer from hazardous to non-toxic texts are presented here, specifically for detoxification purposes. Using the additional style guidelines, both of them include high-quality pre-trained LMs. Paraphrasers trained using the style-conditioned GPT-2 model form the basis of Par aGeDi. The BERT-based CondBERT model does not need fine-tuning, and a pre-trained toxicity classifier handles all style control. We undertake a comprehensive investigation of style transfer models that use both automated and human-evaluation methods. Compared to existing cutting-edge style transfer models, our suggested techniques perform better on the detoxification and sentiment transfer tasks. This property is no longer used in the Ethical Statement. To begin with, Unexpected outcomes and research byproducts might cause damage while discussing toxicity, a delicate subject. So, let's think about some ethical issues that are connected to our job. Regarding Toxicology Almost any kind of bad behavior on the Internet may be described as toxic. It may be as subtle as using a condescending tone (Perez Almendros et al., 2020) or as egregious as engaging in racial or socioeconomic profiling as a basis for oppression.

Although there is consensus among annotators when it comes to identifying instances of extreme toxicity, such hate speech, the classification of less severe forms of toxicity is subjective and influenced by the annotator's background (Al Kuwatly et al., 2020; Fortuna and Nunes, 2018). Because of this, the severity of certain forms of poisoning may be underestimated. We explicitly offer a data-driven strategy in Appendix A to define toxicity in the most objective and possible manner. Our two proposed models can detect toxicity using a toxicity-labeled dataset without the need for any extra rules or dictionaries that have to be constructed by hand. As a result, they may adjust their toxicity perception based on the facts they enter. Given a corpus with objective toxicity labels, our models will be able to deliver impartial detoxification results. However, it should be used with care since the model might replicate biases in the training corpus. Degradation of Written Works It is possible to rework a neutral text into a harmful one, which is the inverse of what a detoxification job entails. Our style transfer approach is only one of several that may theoretically do this. On the other hand, concocting the outcomes of this "toxification" into actual poisonous phrases would be almost to impossible with CondBERT because to the poor quality of such trans creation. Toxic data structures are to blame for it. Lexical indicators of toxic style, such as unpleasant or profane terms, are one of its primary characteristics. The presence of these markers strongly indicates that they belong to this class since they (i) convey the majority of a sentence's stylistic information and (ii) have synonyms that do not. We depend heavily on these qualities for both of our strategies. By doing so, they are able to substitute harmful words with safer alternatives. By contrast, there are no non-toxic terms that strongly indicate a neutral (non-toxic) style if we do the inverse change. The second issue is that it is very difficult to find non-toxic terms with toxic synonyms and find suitable replacements. Thus, we propose that CondBERT should not be used for detoxification. None of the points presented above disprove the toxificative potential of Cond BERTor ParaGeDi. On the other hand, they imply that simpler toxification strategies (such handwritten instructions for adding nasty words) can produce lower-quality literature. The Censorship of Detoxification An other cause for worry is the potential for detoxifying technology to be used for the purpose of censorship by altering user-generated content. Here, we'd like to take a step back and examine things from a new angle. Censorship is already a part of social media; for instance, Instagram has mechanisms to remove messages that include hazardous information that it detects automatically. 5 Conversely, we propose a strategy to lessen the impact of this regulation by rephrasing harmful signals rather than removing them completely. Finally, we recommend that user messages cannot be changed without the user's explicit agreement. It is more appropriate to utilize the detoxification models to propose detoxifying modifications than to automatically implement them. Simultaneously, detoxification models have the potential to enhance chatbot security by purifying (if required) responses before to transmission to consumers. One unresolved issue with neural chatbots is that they may be pre-trained on biased textual input, which can lead to automatically produced harmful comments (Gehman et al., 2020). According to Babakov et al. (2021), a potential use-case for avoiding reputational damage for the organization that produced the unmoderated chatbot might be to cleanse autonomously generated material.

References

1. Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 184–190, Online. Associa tion for Computational Linguistics.
2. Nikolay Arefyev, Boris Sheludko, Alexander Podol skiy, and Alexander Panchenko. 2020. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitu tion. In Proceedings of the 28th International Con ference on Computational Linguistics, pages 1242 1255, Barcelona, Spain (Online). International Com mittee on Computational Linguistics.
3. Nikolay Babakov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Detecting inappropriate messages on sensitive topics that could harm a company's reputation. In Proceed ings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 26–36, Kiyv, Ukraine. Association for Computational Linguistics.
4. Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: Asimple approach to controlled text generation. In 8th International Conference on Learning Represen tations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under standing. In Proceedings of the 2019 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Associ ation for Computational Linguistics.

6. Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In Proceed ings of the Second Workshop on Trolling, Aggres sion and Cyberbullying, pages 21–25, Marseille, France. European Language Resources Association (ELRA).
7. Paula Fortuna and Sérgio Nunes. 2018. A survey on au tomatic detection of hate speech in text. ACM Com put. Surv., 51(4).
8. Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxi cityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.
9. Xiaochuang Han and Yulia Tsvetkov. 2020. Fortify ing toxic speech detectors against veiled toxicity. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7732–7739, Online. Association for Computa tional Linguistics.
10. Junxian He, Xinyi Wang, Graham Neubig, and Tay lor Berg-Kirkpatrick. 2020. A probabilistic formu lation of unsupervised text style transfer. In 8th International Conference on Learning Representa tions, ICLR 2020, Addis Ababa, Ethiopia, April 26 30, 2020. OpenReview.net.
11. Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward con trolled generation of text. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learn ing Research, pages 1587–1596. PMLR.
12. Jigsaw. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment classification-challenge. Accessed: 2021-03-01.
13. Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/c/jigsaw unintended-bias-in-toxicity-classification. cessed: 2021-03-01. Ac
14. Jigsaw. 2020. Jigsaw multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw multilingual-toxic-comment-classification. cessed: 2021-03-01. Ac
15. Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In Pro ceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Pa pers, pages 424–434. Association for Computational Linguistics.