DiffuDetox: A Mixed Diffusion Model for Text Detoxification

Griffin Floto[1], Mohammad Mahdi Abdollah Pour[1], Parsa Farinneya[1], Zhenwei Tang[1], Ali Pesaranghader[2], Manasa Bharadwaj[2], and Scott Sanner[1]
[1] University of Toronto, Canada
{griffin.floto,m.abdollahpour,parsa.farinneya,zhenwei.tang}@mail.utoronto.ca
ssanner@mie.utoronto.ca
[2] LG Electronics, Toronto AI Lab
{ali.pesaranghader, manasa.bharadwaj}@lge.com

Abstract

Text detoxification is a conditional text generation task aiming to remove offensive content from toxic text. It is highly useful for online forums and social media, where offensive content is frequently encountered. Intuitively, there are diverse ways to detoxify sentences while preserving their meanings, and we can select from detoxified sentences before displaying text to users. Conditional diffusion models are particularly suitable for this task given their demonstrated higher generative diversity than existing conditional text generation models based on language models. Nonetheless, text fluency declines when they are trained with insufficient data, which is the case for this task. In this work, we propose DiffuDetox[1], a mixed *conditional* and *unconditional* diffusion model for text detoxification. The conditional model takes toxic text as the condition and reduces its toxicity, yielding a diverse set of detoxified sentences. The unconditional model is trained to recover the input text, which allows the introduction of additional fluent text for training and thus ensures text fluency. Extensive experimental results and in-depth analysis demonstrate the effectiveness of our proposed DiffuDetox.

1 Introduction

Toxic texts with offensive and abusive words are frequently encountered in online forums and social media. Such a harmful online environment can lead to mental health problems (Viner et al., 2019; Wijesiriwardene et al., 2020), which motivates considerable research efforts (dos Santos et al., 2018; Laugier et al., 2021; Logacheva et al., 2022) in text detoxification, i.e., a conditional text generation task aiming to remove offensive content from sentences while preserving their meanings.

Intuitively, there exist diverse ways to detoxify a given sentence. As shown in Table 1, some detoxified sentences are the results of simply removing or replacing the toxic word, e.g., Detoxified 1 and 2, which may cause loss of information or lower text fluency. While other candidates, e.g., Detoxified 3, can reach human-level text detoxification performance with satisfactory fluency and content preservation. Therefore, if a diverse collection of detoxified sentences are given, we can select the most fluent and preservative one to maximize user experience. To do so, we resort to textual conditional diffusion models (Li et al., 2022; Gong et al., 2022) because they are shown to be capable of generating more diverse sets of candidates compared to existing solutions based on transformers (Vaswani et al., 2017), e.g., GPT2 (Radford et al., 2019). Given their demonstrated high generative diversity, diffusion models are particularly suitable for this task.

Toxic	The country doesn't really have to **give a shit** about international laws.
Detoxified 1	The country doesn't really have to give [· · ·] about international laws.
Detoxified 2	The country doesn't really have **care** about international laws.
Detoxified 3	The country doesn't really **need to care** about international laws.
Human	The country doesn't need to care about international laws.

Table 1: A diverse collection of detoxified sentences helps to approach human-level text detoxification.

Nevertheless, previous textual conditional diffusion models (Li et al., 2022; Gong et al., 2022) are not directly applicable to text detoxification due to the scarcity of text detoxification data. Given that text detoxification is a relatively new field and the high cost of human annotations, the available text detoxification data is on the order of $1e^{-1}$ to $1e^{-2}$ of datasets used for other tasks with textual conditional diffusion models (Gong et al., 2022).

To this end, we introduce DiffuDetox, a mixed *conditional* and *unconditional* diffusion model for text detoxification. In particular, the conditional

[1] https://github.com/D3Mlab/diffu-detox

Findings of the Association for Computational Linguistics: ACL 2023, pages 7566–7574
July 9-14, 2023 ©2023 Association for Computational Linguistics

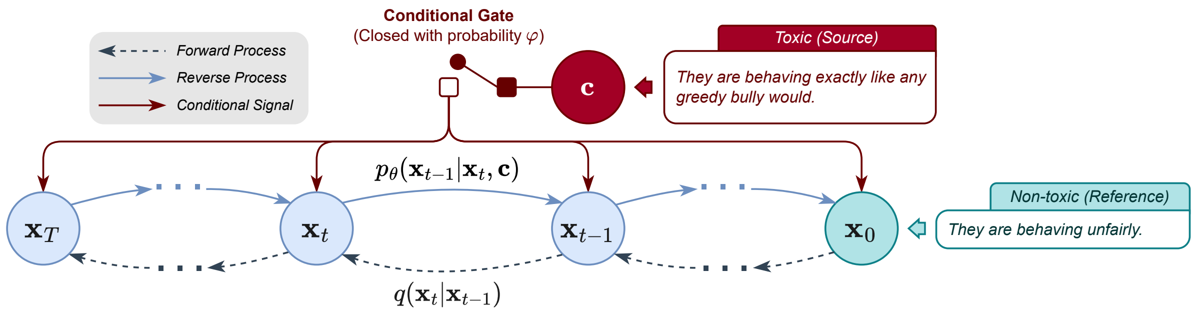

Figure 1: The overall framework of DiffuDetox, a mixed conditional and unconditional diffusion model. For the conditional learning phase, the condition gate is closed with probability φ, then $\mathbf{x}_0$ and $\mathbf{c}$ are sampled from the detoxification dataset. $\mathbf{x}_0$ and $\mathbf{c}$ are set as non-toxic text and toxic text, respectively. For the unconditional learning phase, the condition gate is open with probability $1 - \varphi$, and $\mathbf{x}_0$ is sampled from the fluent text corpus.

model takes toxic text as a condition and through a Markov chain of diffusion steps, yields a diverse set of detoxified sentences. On the other hand, the unconditional model is trained to recover any given input text exactly. That allows us to introduce additional fluent text to be reconstructed by the unconditional model, which is used to improve the fluency of the conditionally generated detoxified sentences. In this way, the resulting diffusion model can maintain a diverse collection of detoxified candidates with satisfactory sentence fluency and content preservation. Extensive experimental results and in-depth discussions demonstrate the effectiveness of DiffuDetox for text detoxification. Our main contributions are summarized in two folds: 1) To the best of our knowledge, we are the first to approach text detoxification with diffusion models, which can maintain a rich collection of detoxified sentences by their high generative diversity; 2) We propose a mixed diffusion model for text detoxification, where the conditional model reduces text toxicity and the unconditional model improves text fluency.

2 Related Work

2.1 Text Detoxification

Previous text detoxification efforts fall into two main categories, *supervised* and *unsupervised*. The unsupervised methods are built on a set of toxic and a set of non-toxic texts without one-to-one mappings between them. Representative methods include Mask&Infill (Wu et al., 2019), DRG-Template/Retrieve (Li et al., 2018), DLSM (He et al., 2020), SST (Lee, 2020), CondBERT and ParaGeDi (Dale et al., 2021). In contrast, the supervised methods are built on parallel datasets in which one-to-one mappings between toxic and non-toxic texts are explicitly provided. ParaDetox (Logacheva et al., 2022) is a well-established method within this category, which fine-tunes BART (Lewis et al., 2020) on their parallel data.

2.2 Textual Diffusion Models

Diffusion probabilistic models are deep generative models with Markov chains of diffusion steps to recover the noise slowly added to data (Sohl-Dickstein et al., 2015). Recently, diffusion models have shown impressive performance on *continuous* domains such as image and audio generation (Ho et al., 2020; Kong et al., 2020), sparking interest in using these models in *discrete* spaces like text. Some textual diffusion models use a discrete diffusion process that operates on word tokens (Savinov et al., 2022; Reid et al., 2022), whereas other methods convert text to embeddings, and then treat text as continuous variables (Li et al., 2022; Strudel et al., 2022). Although textual diffusion models have proved to be effective in various text generation tasks with rich data (Gong et al., 2022), they have not yet been applied to tasks with fewer training samples, such as text detoxification in our case. Ho and Salimans (2021) are the first to exploit unconditional diffusion models for conditional generation, while their method is limited to images and is not aiming for introducing additional data under the low-data setting.

3 Methodology

As the overall framework of DiffuDetox shown in Figure 1 details, our proposed diffusion model for text detoxification improves text fluency in the low-training data regime by using a mixture of a conditional and unconditional diffusion model. We overview diffusion models before discussing DiffuDetox in detail.

3.1 Diffusion Models

Diffusion is a generative modeling paradigm that can be understood as a denoising algorithm (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Song et al., 2021). Noise is gradually added to data samples, while the diffusion model is trained to reverse the process and recover the original data. The framework can be described as a Markov process with T steps, where the original data exist at $t = 0$. Given a sample $\mathbf{x}_0$, the so-called forward process gradually adds noise to the data points, i.e., the blue arrows in Figure 1. The noisy sample can be described by:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x_t}, \beta_t \mathbf{I}) \quad (1)$$

where the variance schedule parameters $\beta_1, \cdots, \beta_T$ are selected such that $\beta_t \in [0, 1]$ and β_0 is close to 0 and β_T is close to 1 (Ho et al., 2020). This ensures that when $t \approx 0$, the data has little noise added to it, while when $t \approx T$, the data is identical to a sample from a standard Gaussian distribution.

The reverse process then attempts to remove the noise that was added in the forward process and is parameterized by θ as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t \mathbf{I}) \quad (2)$$

where the predictive model μ_θ is:

$$\mu_\theta := \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)) \quad (3)$$

which depends on time-dependent coefficients $\alpha := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. In Eq. (3), ϵ_θ is interpreted as predicting the noise that was added to $\mathbf{x}_t$. To optimize the log-likelihood of this model, a simplified training objective is used which reduces the problem to:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2] \quad (4)$$

After training, samples are generated by beginning with pure noise from a standard Gaussian distribution, which is then gradually denoised T times by the learned reverse process.

3.2 DiffuDetox: A Mixed Diffusion Model for Text Detoxification

The task of text detoxification can be viewed as generating a non-toxic sentence, conditioned on a toxic input sentence. The goal is to ensure that the semantics and content of the text are preserved after detoxification, while ensuring that the generated text is fluent. With this interpretation (Gong et al., 2022), we can apply a conditional diffusion model that generated non-toxic text, when conditioned on a toxic sentence. A conditional diffusion model is modified such that the reverse process is now $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$, and the predictive model is $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$. This model can be interpreted as mapping sequences to sequences in a non-autoregressive manner. To apply this model to textual data, sentences are tokenized and converted to a stack of embeddings which are then taken to be $\mathbf{x}_0$ in the diffusion process. When sampling, embeddings that are generated by the diffusion model are converted to tokens by a shallow single-layer decoder.

While diffusion models have high sample diversity which can be used to generate a large number of candidate items, the fluency of the samples is degraded when trained on a smaller dataset. We propose to use a combination of the conditional model diffusion model as well as an unconditional model to tackle this problem. The conditional model is used to detoxify text, whereas the unconditional model can be used to guide the sampling process towards higher quality samples (Ho and Salimans, 2021). The models are combined in a manner that is inspired by the gradient of an implicit classifier $p^i(\mathbf{c}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{c})/p(\mathbf{x})$ such that the following linear combination of the models is used for sampling:

$$\bar{\epsilon}_\theta(\mathbf{x}, \mathbf{c}) = (1 + w)\epsilon_\theta(\mathbf{x}, \mathbf{c}) - w\epsilon_\theta(\mathbf{x}) \quad (5)$$

4 Experiments

4.1 Experimental Settings

Datasets. We conduct our experiments upon a well-established benchmarking dataset ParaDetox[2] (Logacheva et al., 2022), which provides human-annotated one-to-one mappings of toxic and non-toxic sentence pairs from 20,437 paraphrases of 12,610 toxic sentences. We use the same data split of Logacheva et al. (2022) with 671 testing sentences for fair performance comparisons. We further consider the BookCorpus (Zhu et al., 2015),

[2] https://huggingface.co/datasets/SkolkovoInstitute/paradetox

MNLI (Wang et al., 2019), and WikiAuto (Jiang et al., 2020), datasets as additional data for unconditional diffusion model training.

Evaluation Metrics. We follow the well-established text detoxification work (Logacheva et al., 2022) to evaluate DiffuDetox with BLEU, Style Accuracy (STA), Content Preservation (SIM), Fluency (FL), and J score. In particular, STA and FL are computed with pre-trained classifiers (Warstadt et al., 2019) to measure the non-toxicity and fluency of a given sentence, respectively. And we compute SIM using cosine similarity between the input and the generated detoxified text with the model of Wieting et al. (2019). Moreover, we compute J score (Krishna et al., 2020) as the averaged multiplication of STA, SIM, and FL, which is highly correlated with human evaluation as shown by Logacheva et al. (2022).

Implementation Details. We implement our mixed conditional and unconditional models with a single diffusion model where $c = \emptyset$ for the unconditional case. During training, the conditional model is selected with probability $\varphi = 0.8$, and the unconditional model is trained using the non-toxic sentences sampled from the ParaDetox dataset and the additional dataset with equal probabilities. We use the union of the BookCorpus, WikiAuto, and MNLI as the additional dataset. In the test stage, we select the best samples from a candidate set of 20 using the J score. The reported results are from a model trained for $1e^5$ steps with a batch size of 32, and the mixture weighting parameter w in Eq. (5) is set to 5. We use the text detoxification methods listed in Section 2.1 as baselines.

4.2 Experimental Results

Performance Comparison. We have two key observations from the results shown in Table 2. Firstly, our proposed DiffuDetox outperforms most baseline methods on most evaluation metrics, and it is reaching state-of-the-art performance by outperforming ParaDetox on two metrics, demonstrating the effectiveness of our proposed method. Another observation is that DiffuDetox achieves a higher J score than human-level text detoxification. Note that the J score has been shown to be highly correlated with human annotations (Logacheva et al., 2022). This human-level performance of DiffuDetox shows its promise to be deployed in real-world text detoxification scenarios to facilitate users in online forums and social media.

	BLEU	**STA**	**SIM**	**FL**	**J**
Human	100.0	0.96	0.77	0.88	0.66
DRG-Template	53.86	0.90	0.82	0.69	0.51
DRG-Retrieve	4.74	0.97	0.36	0.86	0.31
Mask&Infill	52.47	0.91	0.82	0.63	0.48
CondBERT	42.45	0.98	0.77	0.88	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
ParaDetox	**64.53**	0.89	0.86	**0.89**	***0.68***
Conditional	61.43	0.91	0.87	0.78	0.64
DiffuDetox	62.13	0.92	**0.88**	0.80	*0.67*

Table 2: Text detoxification performance on the ParaDetox dataset. Baseline results are taken from (Logacheva et al., 2022). The best results are in boldface, the strongest baseline performance is underlined, and the J score results reaching human-level detoxification performance are in italics.

Moreover, such results are achieved by selecting from the diverse collection of detoxified sentences generated by diffusion models, which reveals their high generative diversity and the suitability of being applied to text detoxification. Examples of detoxified sentences generated by DiffuDetox can be found in Appendix A.

Ablation Study. We conduct ablations study to investigate the effectiveness of the unconditional model. Since the unconditional model allows the introduction of the additional fluent text, the ablation study can provide insights into the effect of both the unconditional model and the introduced additional data. As shown in Table 2, the model named *Conditional* represents DiffuDetox without the unconditional component. We observe that the addition of the unconditional model improves all the metrics. In particular, text fluency achieves the most significant performance gain. More importantly, the addition of the unconditional model pushes the diffusion model over the human baseline for the J score. Such results demonstrate the effectiveness of the unconditional model and the introduced additional fluent text in improving text fluency and overall performance.

5 Conclusion

In this paper, we approach the text detoxification task with diffusion models for their demonstrated high generative diversity. We introduced DiffuDetox, a mixed conditional and unconditional diffusion model, where the conditional part reduces toxicity whereas the unconditional part ensures fluency. Experimental results show DiffuDetox achieves human-level text detoxification performance, making it promising to be applied in real-world text detoxification systems to benefit users.

Limitations and Future Work

One limitation of our method is that sampling requires sampling both a conditional and a unconditional model, which results in slower inference times. On the other hand, progressive distillation (Meng et al., 2022) provides an attractive solution to this problem. Another limitation is that Ho and Salimans (2021) show that the diversity of generative models is degraded as w increases. Ideally we would be able to have a model that improves upon the fluency as well as the model diversity.

As for future work, we will leverage advanced large language models as the base architecture for training diffusion models to compete with high performance auto-regressive models. Additionally, we will investigate modifications to diffusion models that are inherent to discrete data.

Ethics Statement

Potential Misuse: DiffuDetox can hypothetically be used to obtain toxic sentences from non-toxic sentences. However, the effectiveness of such a scenario should be investigated.

Environmental Cost: We note that while our work required extensive experiments to draw sound conclusions, future work will be able to draw on these insights and need not run as many large-scale comparisons. Models in production may be trained once using the most promising settings.

Acknowledgements

We would like to acknowledge that this work was supported by LG Electronics, Toronto AI Lab Grant Ref No. 2022-1473.

References

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.

Cicero 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.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7943–7960. Association for Computational Linguistics.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis.

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

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.

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.

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

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.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. 2022. On distillation of guided diffusion models.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Machel Reid, Vincent J. Hellendoorn, and Graham Neubig. 2022. Diffuser: Discrete diffusion via edit-based reconstruction.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. 2022. Step-unrolled denoising autoencoders for text generation. In *International Conference on Learning Representations*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. 2022. Self-conditioned embedding diffusion for text generation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Russell M Viner, Aswathikutty Gireesh, Neza Stiglic, Lee D Hudson, Anne-Lise Goddings, Joseph L Ward, and Dasha E Nicholls. 2019. Roles of cyberbullying, sleep, and physical activity in mediating the effects of social media use on mental health and wellbeing among young people in england: a secondary analysis of longitudinal data. *The Lancet Child & Adolescent Health*, 3(10):685–696.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

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

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *International Conference on Social Informatics*, pages 427–439. Springer.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

	Example
Toxic Sentence DiffuDetox Human Reference	that guy used to be a real dick. that guy used to be a real bad man. That guy used to be a mean person.
Toxic Sentence DiffuDetox Human Reference	i didn t say shit! i didn t say anything! I did not say that.
Toxic Sentence DiffuDetox Human Reference	one isn t dead but just totally fucked up. one isn t dead but just totally messed up. One is not dead but just completely destroyed.
Toxic Sentence DiffuDetox Human Reference	kid, you should be watching the crap they call cartoons now made for your generation . kid, you should be watching the stuff they call cartoons now made for your generation. kid, you should be watching the cartoons now made for your generation.
Toxic Sentence DiffuDetox Human Reference	spoiler russia doesn t give a fuck what canada thinks. spoiler russia doesn t care what canada thinks. Russia doesn't care what Canada thinks.
Toxic Sentence DiffuDetox Human Reference	how the fuck do you support the economy when you can barely support yourself. how do you support the economy when you can barely support yourself. How to support the economy when you can barely support yourself
Toxic Sentence DiffuDetox Human Reference	harper needs to get the fuck out of office. harper needs to get out of office. Harper needs to get out of office
Toxic Sentence DiffuDetox Human Reference	again , give me the name of the store or fuck off, liar. again, give me the name of the store or go away. again, give me the name of the store.
Toxic Sentence DiffuDetox Human Reference	now that is just a fucking dumb thing to say. now that is just a bad thing to say. now that is just a useless thing to say.

Table 3: Examples for performance comparison of DiffuDetox against human reference

A Appendix

Table 3 shows examples of toxic texts with DiffuDetox paraphrases and human references. DiffuDetox is able to achieve human-level paraphrasing performance as evaluated quantitively in Section 4.2.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.