PAN 2024 Multilingual TextDetox: Exploring Different Regimes For Synthetic Data Training For Multilingual Text Detoxification

Notebook for PAN at CLEF 2024

Nikita Sushko[1,*]

[1]Skoltech, Bolshoy Boulevard, 30, p.1, 121205, Moscow, Russian Federation

Abstract

Multilingual text detoxification is a style transfer task of creating neutral versions of toxic texts across multiple languages. In this paper, we use a mix of real and synthetic data to build a multilingual text detoxification model using a parallel corpus of toxic and non-toxic texts in 9 languages. We evaluate models trained on various combinations of the training data and determine the optimal training regime. Our proposed approach, which combines an ensemble model with a toxic word deletion baseline, achieves a top-3 score in automatic evaluations and a top-4 score in manual evaluations in the TextDetox 2024 shared task.

Keywords

PAN 2024, Multilingual Text Detoxification (TextDetox) 2024, style transfer, multilingual detoxification, text generation, evaluation, competition, metrics analysis, crosslanguage knowledge transfer, synthetic data

1. Introduction

The proliferation of online social networks has given rise to new challenges in maintaining safe and respectful digital environments. With the increasing prevalence of toxic language, such as hate speech and profanity, online communities face significant threats to their well-being and cohesion. In response, some social media platforms like VK have implemented measures to classify user-generated content as "toxic" or "non-toxic," offering users alternative responses, like stickers or emojis, to convey their intended meaning without resorting to offensive language. However, these approaches are limited in their ability to address the broader issue of toxic content, since users can simply ignore these suggested stickers and send toxic messages anyway.

One promising approach to this problem is text detoxification – a technique aimed at transforming potentially offensive input into neutral output without compromising its original meaning or intent.

In this paper, we propose a two-stage algorithm for multilingual text detoxification, using a finetuned bigscience/mt0-xl[1] [1] model on a mix of publicly available and synthetic data and deletion of toxic words. The pipeline of synthetic data generation is also presented. Different training regimes with various mixes of synthetic and real data are explored and optimal training regime is determined. The resulting synthetic dataset[2] and detoxification model[3] are available on HuggingFace.

During automatic evaluation, the resulting algorithm achieved third place across all languages and fourth place during manual evaluation in the PAN at CLEF Multilingual Text Detoxification (TextDetox) 2024 shared task [2, 3].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

nikita.sushko@skoltech.ru (N. Sushko)

[1]bigscience/mt0-xl on HuggingFace https://huggingface.co/bigscience/mt0-xl

[2]chameleon-lizard/synthetic-multilingual-paradetox on HuggingFace https://huggingface.co/datasets/chameleon-lizard/synthetic-multilingual-paradetox

[3]chameleon-lizard/detox-mt0-xl on HuggingFace https://huggingface.co/chameleon-lizard/detox-mt0-xl

2. Previous work

Text detoxification is a relatively new field, which started from a paper by dos Santos et. al. [4], where they utilized an encoder-decoder translation model, trained with cycle consistency loss to solve the task of unsupervised detoxification.

More recently, Laugier et. al. [5] proposed finetuning T5 model [6] on a detoxification task, using denoising and cyclic autoencoder loss. In RUSSE-2022 shared task [7] further explored non-English detoxification, with solutions ranging from using decoder-only networks with right prompts to fine-tuning an encoder-only tagger for toxic words and a style transfer encoder-decoder model for further detoxification [8].

In addition to these approaches, a paper by Dale et. al. [9], two algorithms were proposed. CondBERT approach, inspired by Wu et. al. [10], utilized a finetuned BERT model for replacing toxic tokens in the sequence to non-toxic. The second approach, ParaGedi, reframes text detoxification problem as a paraphrase and imposes constraints on toxic tokens used during the generation.

Authors of [11] proposed finetuning a multilingual mBART model on a big parallel corpus of English and Russian texts. Their work has shown, that reformulating the task of detoxification as a neural machine translation task boosts performance of the models, given enough data, outperforming CondBERT baseline. Also, they've proved that finetuning a pretrained multilingual model on any of the languages it knows, not on the main language of the model, is possible.

3. Data

TextDetox 2024 shared task consisted of two phases. During the dev phase of the task, organizers provided a training set consisting of 1000 parallel toxic and neutral samples in Russian and English languages. During the test phase, organizers provided a training set, consisting of 400 parallel toxic and neutral samples in 9 languages: English, German, Spanish, Amharic, Arabic, Hindi, Chinese, Ukrainian and Russian.

Additionally, a non-parallel set of 2500 toxic and 2500 neutral sentences in the same 9 languages was provided, as well as a dataset of toxic lexicon, consisting of swear words in these languages.

3.1. Metrics

To assess the resulting models and given data, we calculated STA, SIM, chrF_1 and J metrics. STA metric measured the style transfer quality using the textdetox/xlmr-large-toxicity-classifier[4] [12] model. The SIM metric can be calculated by finding cosine similarity between the embeddings of sentence-transformers/LaBSE[5] [13] model for the input and output (i.e., toxic and neutral sentences). chrF_1 [14] measures the similarity between model output and the references by using character n-grams. J metric is a multiplication of STA, SIM and chrF_1 metrics. The calculation of metrics was conducted using the evaluation script, provided by competition organizers, with toxic examples as the input and neutral examples as both references and output.

3.2. Data preprocessing

Upon examining the provided data, we found out that it's quality varied significantly from language to language.

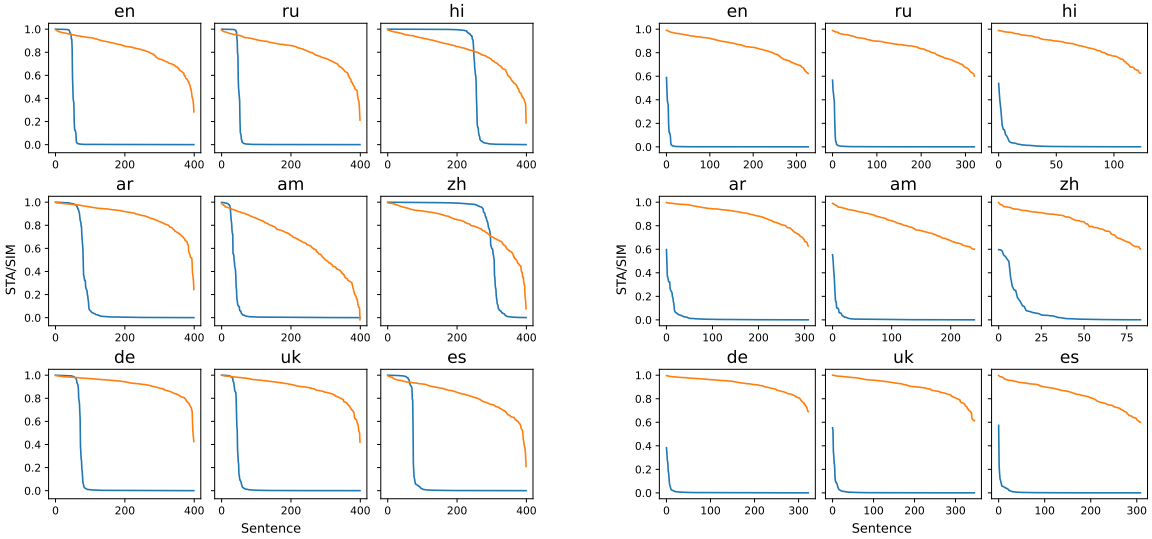
As shown in the Table 1, the quality of provided examples is suboptimal in Chinese and Hindi, as the "neutral" sentences have extremely low STA score. This indicates that only 25% of Chinese and 36% of Hindi neutral examples are actually non-toxic.

Furthermore, the neutral sentences in Amharic language are quite distinct from the toxic sentences, as evidenced by the SIM metric of 0.67.

[4]textdetox/xlmr-large-toxicity-classifier on HuggingFace https://huggingface.co/textdetox/xlmr-large-toxicity-classifier
[5]sentence-transformers/LaBSE on HuggingFace https://huggingface.co/sentence-transformers/LaBSE

Table 1
STA, SIM and amount of real pairs in dirty and cleaned form

Language	STA dirty	STA clean	SIM dirty	SIM clean	Pairs dirty	Pairs clean
en	0.87	0.99	0.82	0.85	400	328
ru	0.87	0.99	0.81	0.84	400	321
uk	0.88	0.99	0.89	0.90	400	347
de	0.82	0.99	0.92	0.92	400	323
es	0.81	0.99	0.82	0.84	400	309
am	0.90	0.98	0.67	0.80	400	241
zh	0.25	0.92	0.80	0.83	400	84
ar	0.79	0.98	0.88	0.89	400	309
hi	0.36	0.98	0.81	0.86	400	124



(a) STA and SIM of real data

(b) STA and SIM of the real data after cleaning

Figure 1: Blue line depicts STA metric. Orange line depicts SIM metric.

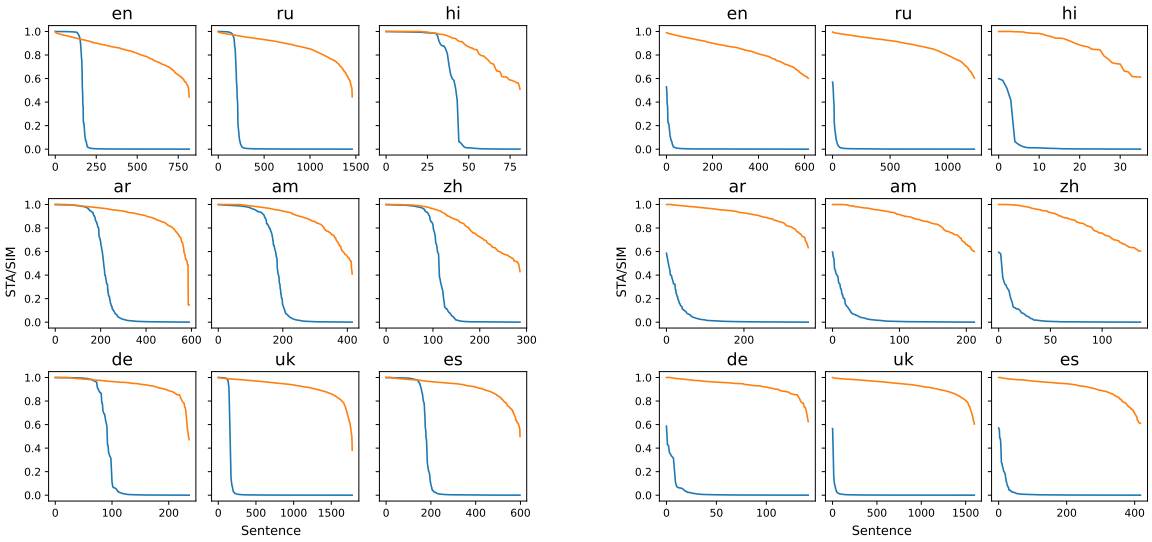
To visualize this, we can sort the sentences in each language by their toxicity scores and plot STA and SIM scores on a graph by a language (Fig. 1a).

By applying a hard threshold of 0.6 for both similarity and toxicity metrics, we can filter out noisy data. However, this approach leads to a drastic reduction in the quantity of examples in Chinese and Hindi languages, with Chinese being left with only 84 examples and Hindi with 120 examples. In addition to that, we can also drop all examples longer than 512 symbols to ensure training stability (Fig. 1b).

3.3. Generating synthetic data

Due to the limited amount of data available after removing non-detoxified pairs from the training data, we need to generate a new dataset. To achieve this, we employed the following algorithm:

1. Train a detox model on uncleaned dataset;
2. Run inference of this model on the toxic sentences from unpaired multilingual dataset;
3. Check if the toxicity classification model classifies the output as non-toxic, if the output of detoxification is still toxic, delete all toxic words from the data, using the toxic lexicon dataset;
4. Check if the toxicity classification model classifies the output as non-toxic;
    - If the output is toxic, do not add the sentence to the resulting dataset;
    - If the output is not toxic, add the sentence to the dataset.



(a) STA and SIM of synthetic data

(b) STA and SIM of the synthetic data after cleaning

Figure 2: Blue line depicts STA metric. Orange line depicts SIM metric.

Table 2
STA, SIM and amount of synthetic pairs in dirty and cleaned form

Language	STA dirty	STA clean	SIM dirty	SIM clean	Pairs dirty	Pairs clean
en	0.79	0.99	0.82	0.84	818	617
ru	0.86	0.99	0.87	0.88	1461	1233
uk	0.91	0.99	0.92	0.92	1778	1599
de	0.61	0.97	0.94	0.93	237	143
es	0.70	0.98	0.91	0.91	599	418
am	0.57	0.95	0.89	0.88	416	213
zh	0.61	0.95	0.83	0.85	287	138
ar	0.64	0.95	0.90	0.91	591	367
hi	0.50	0.93	0.87	0.87	82	36

For the toxicity classification model, we utilized the intfloat/multilingual-e5-large model [15], which was trained on non-parallel data with an 80/20 train-test split. In contrast, for the detox model, we employed the bigscience/mt0-xl model [1]. We trained it for one epoch on all languages, using the AdamW optimizer with a learning rate of 1e-4, a constant scheduler, and a batch size of 6. All training was performed in full precision. The rationale behind choosing this model and its evaluation are presented in the Experiments section.

Although the resulting dataset is of lower quality than the real dataset, after applying the same cleaning procedure, its metrics become comparable to those of the cleaned original dataset (Fig. 2a, 2b, Table 2). By combining these two datasets, we obtained the training data for the final model.

4. Experiments

4.1. Motivation for choosing the model

There are various approaches to tackle the problem of text detoxification. One possible method is to employ encoder-only models, such as BERT, to identify toxic words in a sentence, mask them, and then treat the problem as a denoising task. However, given that we have a dataset consisting of parallel data (i.e., toxic and neutral versions of the same sentence), it is more intuitive to view this problem as a sequence-to-sequence task. Therefore, selecting a full transformer model is the obvious choice for this

Table 3
Evaluation metrics of models trained on different data types: Dirty Real (original competition data before cleaning), Dirty Synthetic (generated data before cleaning), Clean Real (competition data after cleaning), and Clean Synthetic (generated data after cleaning). Cleaning was done with pipeline, which is explained in 3.2. Eval data is a 10% random sample of dirty real data. Best results are in **bold**.

Regime	STA	SIM	chrF_1	J
Dirty Real	0.64	0.89	0.70	0.41
Dirty Synth	0.69	0.88	0.65	0.41
Dirty Real + Synth	0.68	0.85	0.66	0.41
Dirty Synth + Real	0.68	0.90	0.69	0.43
Dirty Mixed	0.7	**0.92**	0.69	0.44
Cleaned Real	0.71	0.90	0.72	0.477
Cleaned Synth	0.71	0.90	0.66	0.437
Cleaned Real + Synth	0.72	0.87	0.71	0.44
Cleaned Synth + Real	0.73	0.88	0.68	0.454
Cleaned Mixed	**0.74**	0.89	**0.73**	**0.481**

problem.

There are three primary families of multilingual encoder-decoder transformer models: mT5, UMT5, and mT0. mT5 [1] is a T5-like model [6] trained on multilingual data. UMT5 [16], on the other hand, shares the same architecture as mT5 but utilizes a novel language sampling algorithm for better dataset creation. It has been demonstrated that UMT5 models outperform mT5 models of the same size across a wide range of tasks. mT0 [1], meanwhile, involves fine-tuning mT5 models on an instruction set, similar to FLAN-T5 [17].

Our experiments show that fine-tuned mT0 models perform better in the task of text detoxification, which led us to adopt the mT0 family as the foundation of our detoxification pipeline. Specifically, we opted for the bigscience/mt0-xl[6] [1] model, as it was the largest model that could fit on our GPU without relying on techniques like LoRA [18]. In addition to mT0-xl, we explored the use of mT5-xl[7] and aya-101 models [19][8]. However, mT5-xl underperformed due to the lack of instruction tuning, while the aya-101 model was too large to be trained on our GPU. We also attempted to utilize LoRA for this task, but even using high rank hyperparameter, the resulting model's performance remained inferior to that of the selected mT0-xl model.

4.2. Exploring different synthetic data training regimes

During training, we explored ten different approaches to training models on synthetic data. We examined training models on real and synthetic data before and after cleaning, mixing the synthetic and real data before and after cleaning, and sequentially training on real + synthetic and synthetic + real data in a two-stage fashion, both before and after cleaning.

The models were trained using the following parameters: AdamW optimizer [20], inverse square root scheduler, learning rate (lr) = 8e-5, batch size (bs) = 4. The training was done in full precision. The models were trained for one epoch.

The best-performing model, according to evaluation set metrics (Table 3), is the model trained on a mix of synthetic and real data. We attribute this to the fact that adding synthetic data to the mix increases the STA metric, which is the hardest metric to optimize. Given enough training steps, the model learns more toxic words and becomes better at deleting them from the input data. Additionally, it is interesting to note that training on synthetic data boosts the STA metric and lowers the chrF_1 metric.

[6]bigscience/mt0-xl on HuggingFace https://huggingface.co/bigscience/mt0-xl
[7]google/mt5-xl on HuggingFace https://huggingface.co/google/mt5-xl
[8]CohereForAI/aya-101 on HuggingFace https://huggingface.co/CohereForAI/aya-101

Table 4
First 5 results after automatic evaluation. The leaderboard is based on J metric. Top-3 best results are highlighted with **bold**. Top-1 result is both **bold and underlined**.

User	average	en	es	de	zh	ar	hi	uk	ru	am
adugeen	0.523	**0.602**	**0.562**	**0.678**	**0.178**	**0.626**	**0.355**	**0.692**	**0.634**	**0.378**
lmeribal	0.515	**0.593**	**0.555**	**0.669**	0.165	**0.617**	**0.352**	**0.686**	**0.628**	**0.374**
nikita.sushko	0.465	**0.553**	0.480	**0.592**	**0.176**	**0.575**	0.241	**0.668**	**0.570**	**0.328**
VitalyProtasov	0.445	0.531	0.472	0.502	0.175	0.523	0.320	0.629	0.542	0.311
erehulka	0.435	0.543	0.497	0.575	0.160	0.536	0.185	0.602	0.529	0.287

Two stage training yields middling results in both chrF_1 and STA, providing better scores than the worst models. The mixed training regime comes out on top, boasting both higher STA and chrF_1 then all other training regimes, although with slightly reduced SIM scores.

Cleaning the data significantly boosts both chrF_1 and STA metrics and moderately improves the SIM metric. The model trained on cleaned version of the real data, outperforms all models trained on non-cleaned data, even when we mix in the synthetic data.

Thus, the optimal approach for training detoxification models in this particular setting is to utilize the Cleaned Mixed training regime, which involves cleaning both synthetic and real datasets from the pairs where neutral outputs are still toxic or where the toxic and neutral sentences are dissimilar, and then mixing them together into one large training set on which the model is trained.

4.3. Final model training

The bigscience/mt0-xl[9] model, trained on a mix of synthetic and real data, was used for the final submission. The training parameters were as follows: AdamW optimizer, inverse square root scheduler, a learning rate of 8e-5, a batch size of 4. The model was trained in full precision for two epochs.

To ensure the model generated responses in the correct language, we used the following prompt: *"Write a non-toxic version of the following text in 'language': 'toxic sentence'."* Without this prompt, the model tended to respond in a language different from the input. The final submission was based on a combination of answers from different models, taken from different training checkpoints.

Notably, the models sometimes failed to detoxify sentences and left out words that could be deleted simply by cutting them out. To address this, each output in the submission pipeline was additionally detoxed using the "delete" baseline method.

5. Results

Our final model achieved third place in the automatic evaluation and fourth place in the manual human evaluation.

During the automatic evaluation, our model consistently ranked within the top three (Table 4), only being outperformed by other models in Spanish and Hindi. The model visibly struggled with scores on Chinese and Hindi datasets, where it performed much worse then in other languages. The reason behind this is that provided data after cleaning was insufficient for training a quality detoxification model and we had to rely on delete baseline for detoxification on Chinese language. We have tried to mitigate it by providing it synthetic data, but after cleaning it from non-detoxified samples, the amount of data was still insufficient for training a good detoxification model on these languages.

In the human evaluation, our model secured first place in Arabic detoxification and ranked among the top three models in Arabic, German, and Hindi (Table 5). Notably, our model outperformed human evaluators in Arabic and German languages in the human evaluation subset. You can see some examples of detoxification in the Table 6.

[9] bigscience/mt0-xl on HuggingFace https://huggingface.co/bigscience/mt0-xl

Table 5
First 5 results after manual evaluation. The leaderboard is based on J metric. Top-3 best results are highlighted with **bold**. Top-1 result is both **bold and underlined**.

User	average	en	es	de	zh	ar	hi	uk	ru	am
Human References	0.85	0.88	0.79	0.71	0.93	0.82	0.97	0.90	0.80	0.85
SomethingAwful	0.77	0.86	**0.83**	**0.89**	0.53	0.74	**0.86**	**0.69**	**0.84**	**0.71**
adugeen	0.74	0.83	0.73	0.70	0.60	**0.82**	0.68	**0.84**	**0.76**	**0.71**
VitalyProtasov	0.72	0.69	**0.81**	0.77	0.49	**0.79**	**0.87**	0.67	0.73	0.68
nikita.sushko	0.71	0.70	0.62	**0.79**	0.47	**0.89**	**0.84**	0.67	0.74	0.68
erehulka	0.71	0.88	0.71	**0.85**	**0.68**	0.78	0.52	0.63	0.65	0.69

Table 6
Examples of detoxification

Language	Toxic input	Neutral output
English	almost as fucked up as the cia funding and arming bin laden.	Almost as bad as the cia funding and arming bin laden.
German	Die wollen von nix ablenken. Die sind naturgemäß so dumm.	Die wollen von nix ablenken. Die sind naturgemäß so naiv.
Spanish	Que mierda es esa ubicación millenial ahora estoy re enojado loko	Que mala es esa ubicación ahora estoy re enojado
Ukranian	Блять, уйобую нахуй в Острог і не вертаюсь. впізду	Уже йду в Острог і не вертаюсь.
Russian	дело даже не в iq - просто хуесосам там не место	дело даже не в iq - просто плохим людям там не место

6. Conclusions

In conclusion, we have identified the optimal approach to training a multilingual seq2seq model for text detoxification tasks. When combined with the detoxification via toxic word deletion baseline, our resulting model achieved third place in the automatic evaluation stage of the PAN 2024 TextDetox competition. The model and dataset are available for download on HuggingFace.

References

[1] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, 2023. arXiv:2211.01786.

[2] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[3] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the

CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.
[4] C. N. dos Santos, I. Melnyk, I. Padhi, Fighting offensive language on social media with unsupervised text style transfer, 2018. arXiv:1805.07685.
[5] L. Laugier, J. Pavlopoulos, J. Sorensen, L. Dixon, Civil rephrases of toxic texts with self-supervised transformers, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), 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. doi:10.18653/v1/2021.eacl-main.124.
[6] 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, 2023. arXiv:1910.10683.
[7] D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, Russe-2022: Findings of the first russian detoxification shared task based on parallel corpora, 2022, pp. 114–131. doi:10.28995/2075-7182-2022-21-114-131.
[8] I. Gusev, Russian texts detoxification with levenshtein editing, 2022. arXiv:2204.13638.
[9] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, 2021. arXiv:2109.08914.
[10] X. Wu, S. Lv, L. Zang, J. Han, S. Hu, Conditional bert contextual augmentation, in: J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, J. J. Dongarra, P. M. Sloot (Eds.), Computational Science – ICCS 2019, Springer International Publishing, Cham, 2019, pp. 84–95.
[11] D. Moskovskiy, D. Dementieva, A. Panchenko, Exploring cross-lingual text detoxification with large multilingual language models., in: S. Louvan, A. Madotto, B. Madureira (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 346–354. URL: https://aclanthology.org/2022.acl-srw.26. doi:10.18653/v1/2022.acl-srw.26.
[12] textdetox, xlmr-large-toxicity-classifier model on huggingface, https://huggingface.co/textdetox/xlmr-large-toxicity-classifier, 2024. Accessed: 2024-05-15.
[13] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, 2022. arXiv:2007.01852.
[14] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/W15-3049.
[15] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).
[16] H. W. Chung, X. Garcia, A. Roberts, Y. Tay, O. Firat, S. Narang, N. Constant, Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=kXwdL1cWOAi.
[17] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.
[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.
[19] A. Üstün, V. Aryabumi, Z.-X. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, F. Vargus, P. Blunsom, S. Longpre, N. Muennighoff, M. Fadaee, J. Kreutzer, S. Hooker, Aya model: An instruction finetuned open-access multilingual language model, 2024. URL: https://arxiv.org/abs/2402.07827. arXiv:2402.07827.

[20] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.