Progress Report

Minsi Lai and **Chenxin Wang** and **Jingyi Liao** and **Fangge Liao**
University of British Columbia

Abstract

This document serves as the progress report for the 6-week team project assignment of COLX_531.

1 Timeline

The official timeline of the shared task titled *Multilingual Text Detoxification (TextDetox) 2024* is per the following:

- **February 1, 2024:** First data available and run submission opens
- **May 6, 2024:** Run submission deadline and results out
- **May 31, 2024:** Participants paper submission

The test data release date is not explicitly given in the shared project information page. But we assume that the test data will be released before May 6, which is the run submission deadline and the results will be out. All submissions need to be submitted to tira.io

2 Task Description

Text detoxification is a task that transforms a toxic text into a non-toxic one while preserving the original content's meaning as much as possible. We aim to create detoxification systems for this task, for at least two languages. The systems take toxic comments as input and produce detoxified comments as output, where comments are considered as toxic due to the presence of harmful language, obscenity, or rudeness.

Text detoxification is critically important as social media platforms and online forums have many toxic comments which can lead to negative experiences for users. This task helps create a safer, more inclusive online environment. Moreover, it provides valuable insights into computational linguistics and natural language processing by tackling the complexity of language understanding and the nuances involved in the tone and intent of the text.

The task can be divided into three major steps with the objectives as follow: (1) Style Transfer: Given the generated paraphrase, accurately classify its level of non-toxicity and convert all toxic paraphrases into non-toxic ones. (2) Content Preservation: The transformed non-toxic sentences mean the same as the original toxic sentences. (3) Grammatical Correctness: The transformed non-toxic sentences are grammatical and easy to understand. Together, these steps form the foundation of the evaluation metrics detailed in a later section, assessing the text detoxification process.

3 Data

3.1 High-level Description of the Official Datasets

This study utilizes two official ParaDetox datasets for English and Russian available, used for fine-tuning task. Each dataset comprises Detoxification with Parallel Data, where the input consists of a toxic comment, and the corresponding output is a detoxified version of the comment (neutral comment). Each comment is a single sentence, approximately 20-30 words in length. The size of datasets and specific lengths of comments are detailed in the descriptive statistics section below. It is important to note the absence of a designated validation split within the official English dataset. Consequently, we have allocated 10% of the English training dataset for validation purposes, ensuring consistency by setting a fixed seed.

In the official development dataset, there are nine distinct splits corresponding to nine languages. Given the available train data of English and Russian, only the respective splits from the development dataset are employed. Each language split within the development dataset encompasses 400 toxic comment entries.

Language	Split	Number of samples	Mean length	Max length	Min length
English	train_toxic	17,769	11.85	20	1
English	train_neutral	17,769	9.30	27	1
English	valid_toxic	1,975	12.00	20	5
English	valid_neutral	1,975	9.44	23	1
English	dev_toxic	400	11.96	24	4
Russian	train_toxic	11,090	10.37	28	1
Russian	train_neutral	11,090	8.93	29	1
Russian	valid_toxic	1,116	10.34	20	5
Russian	valid_neutral	1,116	8.66	24	1
Russian	dev_toxic	400	10.49	25	4

Table 1: Descriptive statistics of the official dataset

3.2 Descriptive Statistics of the Official Datasets

Table 1 shows the descriptive statistics of the two official ParaDetox datasets and the official development dataset for English and Russian. For the "Split" feature, "_toxic" means toxic comments (input) and "_neutral" means neutral/detoxified comments (output), e.g. "train_toxic" means inputs in the train dataset.

4 Evaluation metrics

There will be both automatic and manual evaluation for this project, where the automatic evaluation script was provided and already stored in the project repo for future integration. The evaluation metrics consist of three key parameters: Style Transfer Accuracy (STA), Content preservation (SIM), and ChrF. All three metric components return value in the range of [0,1], and the final common metric for scoring will be calculated as the mean of STA*SIM*ChrF per sample.

4.1 Style Transfer Accuracy (STA)

STA measures the level of non-toxicity of the generated paraphrase from the model output. This is a binary toxicity classifier that returns the probability of a sentence being classified as non-toxic (label = 0) or toxic (label = 1). This metric is based on a fine-tuned instance of XLM-RoBERTa model, which is a multilingual version of RoBERTa, a pre-trained transformers model. The model takes a sentence as input, and randomly masks 15% of the words, then takes the entire masked sentence through the model to predict the masked words.

4.2 Content preservation (SIM)

SIM measures the similarity between the original toxic sentence and the model generated paraphrase. This metric is calculated as the cosine similarity between the LaBSE embeddings of the two sentences.

4.3 Fluency (ChrF)

ChrF is the character n-gram F score, which is used to measure the similarity between the generated paraphrase and the human-written detoxified references. Calculation of this metric is based on character n-gram precision and recall enhanced with word n-grams. The tool calculates the F-score averaged on all character and word n-grams, where the default character n-gram order is 6 and word n-gram order is 2. Finally, the arithmetic mean is calculated for n-gram averaging.

5 The background

Our team read and summarized the following 3 papers that are related to this task:

5.1 Methods for Detoxification of Texts for the Russian Language

This paper focuses on the problem of automatic detoxification of Russian texts to combat offensive language, which shed light on textual style transfer that can be used for auto-eliminating toxic text in various contexts (Dementieva et al., 2021). The main task addressed is the detoxification of toxic text content in Russian language, specifically to perform style conversion where the source style is toxic and the target style is neutral/non-toxic. This task is crucial for processing toxic content and eliminating toxicity, especially given the lack of

previous work on detoxification for the Russian language.

The dataset used in this study is the RuToxic dataset, which consists of 163,187 texts from Russian social networks, with 19% toxic and 81% non-toxic texts. This dataset was created by merging two corpora of toxic comments released on Kaggle, with adjustments to unify the labeling schemes across different types of toxicity.

The models introduced and compared in the paper are based on the BERT and GPT-2 architectures. Specifically, the paper tests a BERT-based model for local corrections and a supervised approach using a pretrained GPT-2 model for rewriting texts.

The main contributions of the paper include: (1) Introducing the first study on text detoxification for the Russian language, while most existing studies on this topic focus on English language. (2) Comparing two types of models and providing baselines: a GPT-2 based method for rewriting texts and a BERT-based model for targeted corrections, which demonstrate the feasibility of applying these methods to the detoxification task. (3) Creating a evaluation setup for similar tasks in Russian, and providing benchmarks for future research in automatic text detoxification.

5.2 Paper 2: Exploring Methods for Cross-lingual Text Style Transfer: The Case of Text Detoxification

Main Task: The goal of this research is to transfer detoxification ability to another language for which corpus is not available. In this study, a comprehensive research of conducting cross-lingual transfer between languages with multilingual seq2seq models is presented. Additionally, the paper also aims to explore a method that performs text translation and detoxification simultaneously, and also introduces new automatic detoxification evaluation metrics with higher correlations with human judgments (Dementieva et al., 2023).

Dataset: For training supervised text detoxification models, the study employs ParaDetox for English, including 18777 lines of training data, 988 lines of deviation data, 671 lines of test data. For Russian, a Russian parallel text detoxification dataset is utilized in this study, with 5058 lines of training data, 1000 lines of deviation data, and 1000 lines of test data. Both two datasets were collected through manual production of alternative expressions (detoxicated sentences), content preservation check. and toxicity classification.

Models: Pretrained multilingual models are used in this study, including mT5 (covering 101 languages), mBART (covering 50 languages) as models, M2M100 model (trained for translation tasks between 100 languages), are utilized in the study.

Cross-lingual detoxification transfer methods: 1. Backtranslation: Translate input sentences into the language for which a detoxification model is available. In this approach, Helsinki OPUS-MT is used for English-Russian translation; and Yandex is utilized for Russian-English translation. 2. Training Data Translation: If the training data in one language is available, translate the language into another and use it to train a separate detoxification model for this language. 3. Multitask Learning 4. Adapter Training: Utilizes Opusparcus for paraphrasing and en-ru parts of Open Subtitles, Tatoeba, and news_commentary for translation.

Contributions: 1. Present a comprehensive study of cross-lingual detoxification transfer methods. 2. Pioneers the investigation of simultaneous detoxification and translation, testing baseline approaches. 3. Introduce updated automatic detoxification evaluation matrices which are more correlated to human judgements than previous benchmarks.

5.3 Paper 3: Text Detoxification using Large Pre-trained Neural Models

The main task this paper tries to solve is the elimination of toxicity in text, specifically aiming to transform toxic sentences into non-toxic ones while preserving the original content's meaning. This task falls under the broader category of style transfer in natural language processing, with a specific focus on detoxification to make online interactions more civil and respectful (Dale et al., 2021).

The dataset used for training and testing the style transfer models in this paper is the English data from the first Jigsaw competition. The dataset was prepared by dividing comments labeled as toxic into sentences and classifying each of them with a toxicity classifier. Sentences classified as toxic were used as the toxic part of the dataset, and an equal number of non-toxic sentences were randomly picked from the Jigsaw data to form the neutral part. For testing, 10,000 sentences with the highest toxicity score according to the classifier were used.

The paper introduces two novel models for text detoxification: (1) ParaGeDi (Paraphrasing

GeDi): This model combines a well-performing paraphraser with guidance from style-trained language models to maintain text content and remove toxicity. It is capable of fully regenerating the input by guiding the generation process with small style-conditional language models and using paraphrasing models for style transfer. (2) CondBERT (Conditional BERT): This method uses BERT to replace toxic words with their non-offensive synonyms, making the method more flexible by allowing BERT to replace mask tokens with a variable number of words. It performs pointwise editing to replace toxic spans found in the sentence with non-toxic alternatives.

The main contribution of this paper includes the proposal of two novel detoxification methods based on pre-trained neural language models (ParaGeDi and CondBERT), a large-scale comparative study of style transfer models on the task of toxicity removal, and the creation of an English parallel corpus for the detoxification task by retrieving toxic/safe sentence pairs from the ParaNMT dataset, demonstrating that it can further improve the best-performing models.

6 Week 1 plan

1. **Baseline model:** we plan to follow the baselines provided by the shared task, which include a backtranslation baseline that performs text detoxification using a sequence of translation, detoxification, and backtranslation processes, and a trivial baseline that simply removes all stopwords.

2. **Pretrained model:** we plan to use a BERT-based model named CondBert, which uses BERT to replace toxic words with non-offensive synonyms. Since we are also dealing with Russian language on top of English, we will likely also incorporate backtranslator Helsinki OPUS-MT for English-Russian translation and Yandex for Russian-English translation.

3. **Hardware:** We plan to use GPUs for model development

4. **Reason for choosing model:** BERT-based architecture is commonly used in the studies we read, and the CondBERT model provides flexibility in handling toxic words by providing a variable number of words for replacement, which we think suits this shared task well.

7 Week 2 Baseline - Backtranslation Model

In week 2, we modified the Backtranslation model from the official website of PAN CLEF 2024 as our baseline model.

7.1 Model Description

The preparation step of the baseline to is load three necessary pre-tained model and corresponding tokenizer pairs. This includes:

- **Translator Model** used for translating text between multiple languages. This comprises the facebook/nllb-200-distilled-600M model on Hugging Face, using the "M2M100ForConditionalGeneration" class for the model and "NllbTokenizerFast" for the tokenizer.

- **English Detoxifier Model** used for reducing toxicity in English text. This consists of the s-nlp/bart-base-detox model on Hugging Face, utilizing the "BartForConditionalGeneration" class for the model and "BartTokenizerFast" for the tokenizer.

- **Russian Detoxifier Model** used for reducing toxicity in Russian text. This consists of the s-nlp/ruT5-base-detox model on Hugging Face, utilizing the "T5ForConditionalGeneration" class for the model and "T5TokenizerFast" for the tokenizer.

Then depending on the language of text to process, the baseline model selects an appropriate pretrained model and the corresponding text tokenizer. For English and Russian input texts, the model will directly apply the detoxifier, while for other input language, the model will first translate the input text into English, then apply the detoxifier for English. After detoxification, a translator is applied again to translate the detoxified English text back into the original language.

During this text processing process, the scripts divides the input text into smaller batches, each containing several pieces of text to enhance processing efficiency. The final step of the baseline is to store the output results. The processed texts are saved to a new file specified by users, with each piece of text tagged with the same id as the input file. Additionally, if chosen, the script can log information to Weights & Biases.

7.2 Error Analysis

Keeping the goals in mind that the goal is to remove or alter offensive language, reducing aggressiveness, or making the text more neutral or formal while attempting to preserve the original meaning as much as possible, an example of Successful Detoxification with Minimal Content Change could be:

- Original: "My page should be protected first so that worthless scum like you can't keep vandalizing it."
- Modified: "My page should be protected first so that someone like you can't keep vandalizing it."
- Analysis: The modification successfully removes the offensive term "worthless scum" and replaces it with "someone," achieving detoxification while preserving the original content.

Below are some analyses on a few notable less ideal attempts to identify patterns and potential issues:

Failure to Detect Toxic Language

- Original: "and go on saddam execution talk page, before you stink up the place."
- Modified: "and go on saddam execution talk page, before you stink up the place."
- Analysis: No changes were made, indicating either an error in processing or a determination that no detoxification was needed.

Error in Detection of Toxic Language

- Original: "then all of a sudden i see her , shes now got the big phony tits and everything ."
- Modified: "then all of a sudden i see her, shes now got the big phony tits"
- Analysis: The part of the original sentence that was changed was not in fact toxic and the truly toxic language was left unmodified.

Loss of Original Meaning Original: "Next time, I will force you to eat your own shit." Modified: "Next time, I will force you to eat your own." Analysis: The original expression was meant to convey that people are to be held responsible. While the resulting change did detoxified it, that aspect of meaning was lost.

Grammatical Mistakes Original: "gt today was one of the most fucked up days of my life ." Modified: "Today was one of the most bad days of my life." Analysis: Replacing "fucked up" with "bad" is correct for this task, but considering that "bad" follows "most", they should just be replaced with "worst", which is more grammatically sound.

General Observations:

- Limitations of the Modification Strategy: The strategy is singular — identifying the toxic word and replacing or removing it, which could be challenging to generating fluent and grammatically sound language and in dealing with diverse contexts that requires attention to the meaning it creates with surrounding words.
- Challenges in Preserving Original Sentiment: Removing offensive language without diluting the original sentiment or intensity poses a challenge. In some cases, the modifications might oversimplify or weaken the original statements.
- Error in Consistency and Unnecessary Retention: Some sentences are left unchanged even when they might benefit from slight modifications for detoxification, indicating possible limitations in the detoxification algorithm's sensitivity or specificity.

Recommendations for Improvement: For this baseline model, the limitation is mostly imposed by the modification strategy, but as we move to later more refined models we expect the problems to be lessened greatly.

7.3 Optional

7.3.1 Paper1: ParaDetox: Detoxification with Parallel Data

Main Task: The main task involves suggesting a pipeline for collecting parallel data for detoxification, which includes crowdsourcing tasks for reducing toxicity while preserving content, checking for semantic similarity, and identifying swear words.

Dataset: The pipeline for collecting parallel data includes three tasks. The first crowdsourcing task asks users to reduce toxicity in each sentence while keeping the content. After getting the

generated paraphrases with their original variants, users are asked to check if they have close meanings. Finally, users are asked to indicate whether the paraphrases contain any swear words. Toxic sentences in ParaDetox, a parallel detoxification dataset, are fetched from three sources: Jigsaw dataset of toxic sentences (Jigsaw, 2018), Reddit and Twitter datasets used by Nogueira dos Santos et al. (2018). The authors selected 7000 toxic sentences from each dataset, then paraphrase 12610 toxic sentences (20437 paraphrases). ParaNMT dataset is automatically filtered from sources, then 1,400 toxic-neutral pairs are chosen manually. Each sentence has only one paraphrase. The paraphrases were not gained via a chain of translation models.

Models: The paper uses a fine-tuned transformer-based generation model BART for training, with specific hyperparameters such as a learning rate of 3e-5 and 10,000 training epochs.

Contributions: 1. Introduce an innovative pipeline for collecting parallel data for detoxification tasks. 2. Release two parallel corpora, ParaDetox and ParaNMT, representing the first datasets tailored for detoxification model training. 3. The collected datasets have the potential to enhance the performance of detoxification systems significantly.

7.3.2 Paper2: RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora

Main task: This paper aims to create a new parallel corpus comprising toxic sentences and manually crafted non-toxic paraphrases. It seeks to refine the automatic evaluation setup by incorporating aspects of style transfer and emphasizing the importance of manual evaluation in assessing model quality.

Datasets: The pipeline consists of three tasks: Paraphrase generation, content preservation check, toxicity classification. Toxic sentences are taken from Russian datasets of toxic messages from social media: Odnoklassniki (Belchikov, 2019), Pikabu (Semiletov, 2020), and Twitter (Rubtsova, 2012). The dataset is devided into training data (6948 toxic sentences), development data (800 toxic sentences), test data (875 toxic sentences). For each toxic sentence, 1–3 variants of detoxification are paraphrased.

Models The paper introduces four models: Baseline, a rule-based Delete approach, fine-tuning on the ruT5 model and the continuous prompt tuning approach for ruGPT3 model.

Contributions: 1. Introduce a new parallel corpus consisting of toxic sentences and manually crafted non-toxic paraphrases. 2. Employ an established automatic evaluation setup that aligns with style transfer formulations and encompasses various transfer quality aspects. 3. The first attempt to use crowdsourcing for large-scale manual evaluation of a text generation model.

8 Reflection for Week 1 & Week 2

What do you expect to gain from this shared task?

We expect to gain substantial insights into the complexities of text detoxification across multiple languages, like using neural networks to handle different languages' detoxification. This task not only offers a practical experience in handling and processing multilingual datasets but also allows us to explore advanced NLP techniques and models for content moderation that would be useful in real life, like making sure people can talk online without running into offensive comments.

Does your task go well? What goes well and what doesn't go well? What is the biggest barrier? Do you feel that you are on the right track?

On the positive side, we have made significant progress in understanding the task requirements, setting up a structured timeline, and identifying key resources and datasets. The initial phases, including dataset exploration and baseline model implementation, went smoothly thanks to our team's collaborative effort and effective division of labor. However, we encountered challenges in integrating our models with TIRA and Docker for evaluation as it was a new learning experience for which there was little documentation and was more complicated than previous uploading processes, and it slowed us down. Despite these bumps, we are moving in the right direction.

If you could re-start this task again, what are the good practices you want to keep and what are the not-so-good practices you want to improve?

Reflecting on our approach, if we were to restart this task, we would maintain good practices of very thorough literature review, planning and collaboration, which have been important in guiding our project direction. However, we would aim to improve initial technical setup, particularly our understanding and use of TIRA and Docker, to avoid the delays we experienced. An earlier focus on

these technical aspects could make our evaluation process smoother and allow more time for model refinement.

What is your plan for system development in upcoming weeks?

For the upcoming weeks, our plan for system development is focused and we intend to explore and implement new models that could offer better handling of the multilingual text data and more sophisticated detoxification abilities. We are considering experimenting with cross-lingual transfer learning techniques and advanced transformer-based models that have shown promise in related tasks. We lso plan to refine our data pre- and post-processing pipelines to enhance model performance, so to achieve better content preservation and fluency (metrics of the evaluation). Our goal is to test and refine our approaches based on evaluation feedback, and try to make continuous improvement in its performance.

8.1 Contribution to reflection

- Chenxin Wang: 25%, researched on Docker&TIRA, wrote the first section of reflection journal
- Fangge Liao: 25%, researched on relevant language models, wrote the second section of reflection journal
- Minsi Lai: 25%, researched on relevant language models, wrote the third section of reflection journal
- Jingyi Liao: 25%, researched on Docker&TIRA, wrote the final section of reflection journal

9 Week 3 - Finetuning on Baseline Models

9.1 Model Description

This week, our team plans to fine-tune two pretrained models, BART for English and T5 for Russian, using the available training data. Given the modest size of the training dataset, we anticipate that the improvements in model performance, compared to the baseline, will be moderate.

Fine-tuned English BART: We started by finetuning the T5 model on Russian data for one epoch due to GPU limits, which didn't beat the baseline. Then, we created a second, improved version (finetuned-T5-2) with four epochs, learning rate adjustments, gradient clipping, model checkpointing, and early stopping to enhance performance and tackle overfitting and gradient problems.

Fine-tuned Russian T5: we started off with a 1-epoch basic fine-tuning instance (finetuned-T5-1) on full RU training set due to limited GPU resource, to ensure that all steps from model training to output generation could run smoothly. Since the 1-epoch fine-tuned instance failed to outperform baseline, we then trained another instance (finetuned-T5-2) with 4 epoches, a learning rate scheduler, gradient clipping, model checkpointing, and early stopping. These techniques helped prevent overfitting and gradient explosion, and enhanced model performance.

9.2 Results

In the detailed analysis presented, the enhanced models that underwent four epochs of training and were subjected to an optimized training regimen demonstrated superior performance over the baseline model, as evidenced by Tables 2 and 3. These tables highlight the comparative analysis of the finetuned instances against the baseline models across three distinct metrics as well as a combined metric, J (the average of the three metrics for each sample).

Particularly, the BART model, once finetuned, surpassed the baseline model in two of the three metrics, specifically STA and CHrF. This indicates that the refined BART model is adept at producing outputs that are less toxic and bear a closer resemblance to human-authored, detoxified references. However, it fell short in surpassing the baseline in terms of SIM, suggesting a loss of some original text information during the detoxification process.

Regarding the Russian language, the initial finetuned T5 model (T5-1) slightly outperformed the baseline in the STA and CHrF metrics but fell significantly behind in SIM. Following further training epochs and the application of optimization techniques, a second iteration of the finetuned T5 model (T5-2) not only outperformed the baseline in STA and SIM but also matched the baseline performance in CHrF. Notably, T5-2 recorded a marginally superior J-score compared to the baseline.

In summary, while the finetuned models generally excel over the baseline in producing outputs that are less toxic and more stylistically aligned with human-generated detoxified references, they exhibit limitations in preserving the context and meaning of the original toxic inputs.

Model	STA	SIM	CHrF	J
baseline-BART	0.87209177	0.861526623	0.796657874	0.610376342
finetuned-BART	0.878888071	0.84165932	0.835208897	0.624632126

Table 2: Performance comparison of BART models on the English language detoxification task.

Model	STA	SIM	CHrF	J
baseline-T5	0.887429714	0.818380237	0.75003856	0.55603049
finetuned-T5-1	0.887964606	0.814346023	0.752609993	0.555628747
finetuned-T5-2	0.887429714	0.818380286	0.75003856	0.556030523

Table 3: Performance comparison of T5 models on the Russian language detoxification task.

9.3 Error Analysis

English: Analyzing the newly detoxed English sentences produced by the finetuned-BART model, considering the differences between it and the baseline model, provides insight into the advancements and areas needing improvement in text detoxification techniques.

9.3.1 Analysis of Selected Examples:

1. Improved Toxicity Removal with Minor Content Loss:
    - Original: "Prove my comment wrong and post a video of how fucking tough you are."
    - Modified: "Prove my comment wrong and post a video of how tough you are."
    - Reflection: The removal of "fucking" successfully reduces toxicity, aligning with the finetuned-BART's higher STA and CHrF scores. However, the slight modification does not significantly impact the content's meaning, indicating a nuanced approach to detoxification.
2. Maintaining Contextual Meaning with Simplification:
    - Original: "that stupid water pipe is just a fatal accident waiting to happen."
    - Modified: "that water pipe is just a fatal accident waiting to happen."
    - Reflection: By removing "stupid," the model makes the sentence less aggressive without altering the warning about the water pipe's danger, demonstrating a balance between reducing toxicity and preserving content.
3. Slight Loss in Expressiveness:
    - Original: "i go there for work every few months, and i can assure you it fucking sucks."
    - Modified: "I go there for work every few months and I can assure you it is not good."
    - Reflection: While the modified sentence is clearly less toxic, it also loses the original's strong emotional emphasis, reflecting the trade-off between lowering toxicity and maintaining the original sentiment's intensity.
4. Inconsistent Detection and Modification:
    - Original: "and go on saddam execution talk page, before you stink up the place."
    - Modified: "and go on saddam execution talk page, before you stink up the place."
    - Reflection: No changes were made, indicating either a missed opportunity for detoxification or an overcautious approach to preserve content, highlighting areas for improvement in sensitivity and specificity.

9.3.2 Comparison with Baseline Model:

The finetuned-BART model outperforms the baseline in STA and CHrF, indicating it's better at generating outputs with lower toxicity and higher fluency. However, the lower performance on SIM suggests that while the finetuned model is effective at detoxifying content, it occasionally does so at the expense of content fidelity.

9.3.3 Recommendations for Further Development:

- Enhancing Content Preservation: While reducing toxicity, strategies should also focus

on maintaining the original message's expressiveness and emotional weight.

- Balancing Detoxification Techniques: Employing a combination of methods that include context understanding and sentiment analysis could lead to more nuanced detoxification without overly simplifying or altering the content.
- Incremental Testing and Refinement: Continuous iterations with diverse datasets can help improve the model's sensitivity to various forms of toxicity and content complexity.

9.3.4 Future Development Plans:

Given the insights from this analysis, our future plans include exploring models that better balance toxicity reduction with content preservation, possibly by integrating more advanced natural language understanding features. We also plan to refine our approach based on incremental testing feedback, aiming for a more nuanced and context-aware detoxification process.

Russian: Analyzing the Russian detoxed sentences produced by the finetuned-T5 models, and considering the differences between the baseline and new models as per the provided metrics, offers insights into the evolution of text detoxification techniques for Russian content.

9.3.5 Selected Examples Analysis:

1. Improvement in Offensive Language Removal:
    - Original: "пидоры когда будем дышать чистым воздухом. адмистрация это к вам"("fags, when will we breathe clean air. administration this is to you")
    - Modified: "Когда будем дышать чистым воздухом. Адмистрация это к вам"("When will we breathe clean air. Administration this is to you")
    - Reflection: The modified sentence effectively removes the offensive term without altering the sentence's overall appeal for cleaner air, demonstrating the model's ability to target and remove slurs while preserving the message.
2. Changing Offensive to Neutral with Minor Content Alteration:
    - Original: "твоё очко уходит в пользу форума!"("your asshole benefits the forum!")
    - Modified: "твоё лицо уходит в пользу форума!"("your face benefits the forum!")
    - Reflection: The change from "очко"(a vulgar term for anus) to "лицо"(face) significantly reduces the sentence's vulgarity. Although it introduces a slightly different nuance, it maintains the critical intent of contributing to the forum.
3. Preserving Meaning with Reduced Aggressiveness:
    - Original: "и нехуя здесь орать что нам что то принадлежало"("and no fucking yelling here that something belonged to us")
    - Modified: "И нечего здесь кричать, что нам что то принадлежало"("And there's no need to yell here that something belonged to us")
    - Reflection: By replacing "нехуя"(a crude way to say "nothing") with "нечего"(a more neutral "no need"), the model effectively tones down the aggression without losing the original sentiment.

9.3.6 Metrics and Model Comparison:

The finetuned-T5-2 model shows a nuanced improvement over the baseline T5 model, particularly in STA (Stability) and SIM (Similarity), indicating its strength in generating outputs with lower toxicity while attempting to retain textual fidelity. However, the slight underperformance in SIM for the finetuned-T5-1 model suggests that certain nuances of the original text may be lost, potentially due to overemphasis on detoxification at the expense of content preservation.

9.3.7 Reflections and Recommendations:

- Balancing Act: The challenge remains in balancing the removal of offensive language with the preservation of the original message's intent and emotional tone. This is particularly evident in examples where replacing vulgar terms with neutral ones slightly shifts the sentence's impact or meaning.

- Contextual Understanding: Enhancing models' contextual understanding could lead to more accurate interpretations of phrases that require sensitivity to cultural and linguistic nuances.

- Incremental Adjustments: Considering the minimal differences in performance metrics, incremental adjustments and more targeted training on problematic phrases could yield better outcomes.

9.3.8 Future Directions:

Moving forward, the focus will be on refining the balance between detoxification effectiveness and content fidelity. Exploring deeper linguistic and cultural nuances in the Russian language, experimenting with context-aware models, and employing targeted fine-tuning strategies may enhance the models' ability to maintain the original text's essence while removing offensive language. Additionally, further iterations should consider user feedback and real-world application scenarios to fine-tune the models' sensitivity and specificity.

9.4 Reflection

Our fine-tuned models demonstrated slight improvements over the baseline models. The refined T5 model has exhibited slight enhancements in both SIM and CHRF metrics when compared to the pre-trained version. The fine-tuning progress positively influenced their capacity to cleanse text, although the overall gains were not large. The largest barrier we faced during this fine-tuning process was the limitation imposed by GPU memory. These limitations hindered the extent to which we could refine our models and the size of the dataset we could utilize.

In the coming weeks, our plan involves integrating the ru-GPT3.5 model, designed for Russian language processing, and refining data processing pipelines using parallel processing techniques and advanced data cleaning methods. We'll address computational constraints through memory optimization and explore alternative architectures with reduced resource requirements to enhance efficiency and performance.

10 Week 4 - Implementing GPT-based Model

10.1 Model description

Since finetuning the baseline model did not yield significant improvement on metrics, we proceeded to test a GPT-based structure this week. Since GPT-based structure works for both EN and RU, we will be using this structure for both languages.

GPT-based model for EN: We started of with the English (EN) model, employing the GPT2LMHeadModel from Hugging Face alongside its corresponding GPT2Tokenizer. Following the preprocessing of the English training data to align with the model's input specifications, we fine-tuned the model employing a carefully selected set of training arguments designed to enhance the optimization process. This fine-tuned instance of the model significantly surpasses the baseline in terms of CHrF performance.

GPT-based model for RU: Since GPT-based model is capable of handling RU language, we went through a similar process to train the pre-trained GPT2LMHeadModel with RU training data.

10.2 Result

EN model: As shown in Table 4 below, the GPT-based model outperforms the BART-based baseline model in both STA and CHrF metrics. However, it falls short in the SIM metric, resulting in a higher combined score, J, on the EN validation set. This outcome suggests that the GPT-based model excels at generating outputs that are not only less toxic but also more closely aligned with human-written detoxified gold standards. Nonetheless, similar to the baseline, the GPT-based model struggles to produce outputs that retain the original text's information effectively.

RU model: As illustrated in Table 5 below, the GPT-based model surpasses T5-based models in both STA and CHrF metrics, resulting in an enhanced combined score J on the validation set. In line with observations from the EN model, the GPT-based model appears to discard a significant amount of information from the original toxic text. To address this issue, refining our preprocessing methods may be necessary to retain more of the original text's content while still achieving effective detoxification.

Model	STA	SIM	CHrF	J
baseline-BART	0.87209177	0.861526623	0.796657874	0.610376342
finetuned-BART	0.878888071	0.84165932	0.835208897	0.624632126
GPT-based	0.87379324	0.81336805	0.99439450	0.70804977

Table 4: Performance comparison of GPT and BART-based models on the English language detoxification task.

Model	STA	SIM	CHrF	J
baseline-T5	0.887429714	0.818380237	0.75003856	0.55603049
finetuned-T5-1	0.887964606	0.814346023	0.752609993	0.555628747
finetuned-T5-2	0.887429714	0.818380286	0.75003856	0.556030523
GPT-based	0.89375227	0.77951097	0.86684654	0.60007727

Table 5: Performance comparison of GPT and T5 models on the Russian language detoxification task.

10.3 Error Analysis

The outcome shows varying degrees of success in preserving the original meaning and fluency. Below is an error analysis of some examples, focusing on the effectiveness, content/meaning/fluency preservation, and patterns in the modification.

1. Effective Detoxification with Minor Content Loss:

- Original: 'the pirate bay goes down all the fucking time.'
- Modified: 'The pirate bay goes down all the time.'
- Analysis: This is a successful detoxification. The removal of "fucking" effectively reduces the sentence's toxicity without significantly altering its meaning.

2. Over Modification Leading to Meaning Loss:

- Original: 'thats why im in alaska, far away from toxic freaking parents.'
- Modified: "That's why am in Alaska, far away from my parents."
- Analysis: The term "toxic freaking" was removed to detoxify the sentence, but this also significantly reduces the original statement's emotional weight, potentially altering the perceived relationship dynamics.

3. Unnecessary Modifications and Artefacts:

- Original: 'everyone who has ever voted for her should be permanently disenfranchised and sterilized.'
- Modified: 'each and every individual who has casted a ballot at any point ever for her ought to be forever disappointed and cleaned. the:'
- Analysis: The attempt to detoxify has introduced awkward phrasing ("forever disappointed and cleaned") and an unnecessary trailing "the:". The term "cleaned" is a particularly odd choice, possibly indicating a misunderstanding by the model.

4. Misinterpretation and Erroneous Transformations:

- Original: 'what the fuck is next edible forks?'
- Modified: 'what is the next edible forks?:ifyify::::::ify:'
- Analysis: The model successfully removes the offensive language but introduces nonsensical part like (":ifyify::::::ify:"). This indicates a potential processing error or a misinterpretation of the task.

5. Inconsistent Capitalization and Punctuation:

- Original: 'u know you owe me a fuck right?'
- Modified: 'u know you owe me right'
- Analysis: While effectively detoxified, the transformation neglects proper capitalization ("U" should be capitalized) and punctuation (missing question mark), compromising grammatical correctness.

Other: We did notice that for some detoxed sentences, the whole sentence is mostly lost and left with nothing intelligible, such as in the below example (but there is some confusion about whether

this is the desired outcome because the gold data also shows the result of having most information omitted).

- Original: '1 41 the price of starbucks for the same sugary shit',
- Modified': '1'
- Gold': '1'

10.3.1 General Patterns:

- Content Preservation: While most transformations successfully reduce toxicity, several instances demonstrate a loss of nuance or significant alterations to the original message.
- Grammatical Consistency: The model's outputs occasionally suffer from grammatical inconsistencies, such as erratic capitalization, punctuation, and awkward phrasing, which detract from readability and clarity.
- Artefacts and Misinterpretation: Some modifications introduce artefacts or nonsensical sequences, suggesting potential limitations in the model's processing or understanding of the detoxification task.

10.3.2 Recommendations for Improvement:

- Fine-Tuning for Contextual Nuance: Enhancing the model's sensitivity to the emotional and contextual nuances of language could improve its ability to detoxify text without significant loss of meaning or introduction of errors.
- Post-Processing for Grammatical Correctness: Implementing a post-processing step to correct grammatical inconsistencies and remove artefacts could greatly improve the quality of the outputs.
- Training Data Diversity: Including a broader and more diverse set of training examples could help the model better understand the varied ways in which language can be made less toxic while preserving the original intent and sentiment.

The analysis indicates that while the detoxification process is generally effective at removing offensive language, there is significant room for improvement in maintaining the original text's meaning, emotional weight, and grammatical correctness.

10.4 Reflection

We successfully implemented fine-tuned GPT-2 based models this week for both Russian and English. For both English and Russian texts, the new models showed significant improvements in the CHrF metric, suggesting a better character-level alignment with reference detoxified texts. The SIM scores and the generally stable STA metrics did, however, show some slight declines for both models.

It is clear that both models have significantly improved based on those metrics and the validation dataset evaluation. However, from examining the development dataset results—where incomplete detoxification was noted—that the models might not accurately represent the semantic core of the detoxified content. This implies that on unseen data, the models are not operating at their best.

In the coming week, we hope to further refine the models based on GPT-2 by investigating changes to parameters like top_p, top_k, and temperature. We intend to begin investigating a number of the value sets that are mentioned in the research paper. We also plan to experiment with different approaches to further optimize the model's output, for example, investigating techniques such as post-processing enhancement through heuristic filters or rules could serve as an extra layer of quality control. We will also look into other pre-trained models.

11 Contributions

11.1 Milestone 1

- Minsi Lai: 25%, progress report template setup, Timeline and Evaluation metrics, week 1 plan sections of the report, and summarization of research paper one
- Chenxin Wang: 25%, data inspection code and md file, Task Description and Data sections of the report
- Jingyi Liao: 25%, summarization of research paper two, document Team Contract
- Fangge Liao: 25%, summarization of research paper three, document Team Contract
- Complete Week 1 Plan and discuss Team contract in group meeting during lab

11.2 Milestone 2

The primary objective for milestone 2 was the implementation of the baseline model and subsequent

submission to the official leaderboard. All team members worked collaboratively to adhere to the guidelines provided on the official website, which necessitates the utilization of TIRA and Docker for executing the baseline model and generating prediction outputs. Regrettably, we encountered challenges in effectively employing TIRA and, as a result, were unable to finalize our submission to the official leaderboard this week.

In addition to these endeavors, we made modifications to the baseline and evaluation scripts to sync status and official evaluation metrics to W&B, and executed the baseline model and its evaluation on the baseline outputs directly using python scripts. Reflecting on the workflow of week 2, each team member contributed equitably to milestone 2, with an equal distribution of effort at 25% per member. In addition to our collaborative efforts, other individual contributions were as follows:

- Chenxin Wang: 30%, organize the project repository, and document the model description section in the progress report
- Fangge Liao: 30%, perform an error analysis on a sample output, and document this analysis in the error analysis section of the progress report
- Jingyi Liao: 30%, Optional part
- Minsi Lai: 10%, read through TIRA documentation, attempted Docker submission

11.3 Milestone 3

- Minsi Lai: 20%, participate in discussion of model fine-tuning, write model description and results parts of the weekly report
- Chenxin Wang: 40%, implement scripts to fine-tune baseline models; fine-tune Bart and T5
- Jingyi Liao: 20%, reflection part
- Fangge Liao: 20%, perform error analysis for the two finetuned models developed this week, write the error analysis part of the weekly report

11.4 Milestone 4

- Minsi Lai: 25%, attempted GPT-based structure, wrote model and result sections of the report
- Chenxin Wang: 25%, implemented GPT-based structure on EN and RU dataset, wrote reflection part of the report
- Jingyi Liao: 25%, attempted GPT-based structure
- Fangge Liao: 25%, attempted ParaGeDi, wrote error analysis part of the report

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.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Methods for detoxification of texts for the russian language. *Multimodal Technologies and Interaction*, 5(9).