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 mindthat 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 Lanugage

- 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 Lanugage

- 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.

8 Contributions

8.1 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

8.2 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: organize the project repository, and document the model description section in the progress report
- Fangge Liao: perform an error analysis on a sample output, and document this analysis in the error analysis section of the progress 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).