Detoxification Style Transfer Using BART

Askhat Sametov
Nazarbayev University
askhat.sametov@nu.edu.kz

Abstract

The spread of harmful content in the online communication makes it difficult to establish productive and good digital relationships. The use of pre-trained encoder-decoder based model, which was then fine-tuned on a parallel corpora dataset of toxic and non-toxic (neutral) texts, is an effective approach to automatically detoxify text. This approach makes use of sequence-to-sequence models' capacities to convert harmful language into its neutral counterpart, maintaining the original meaning of the message while eliminating offensive components. This paper goes into detail on dataset preparation, model training and results evaluation.

Results demonstrate that final model achieves significant improvements over baseline methods, offering a promising solution to lessen the effects of online toxicity. This work contributes to the broader field of natural language processing by addressing an urgent need for effective tools to improve online discourse.

1 Introduction

The widespread use of internet platforms has completely changed the way individuals communicate, share information, and express themselves. However, this increased connectivity has also given rise to a significant challenge: the spread of toxic content. Toxicity in online environments, characterized by abusive, offensive, or harmful language, can have negative impacts on individuals and communities. It discourages participation, affects mental health, and can perpetuate hate and discrimination. It is crucial to address this issue for maintaining the health of online communities and supporting principles of respectful dialogues.

The task of text detoxification involves the identification and alteration of toxic text to make it harmless while preserving the underlying message and intent. The subjective assessment of what defines a proper conversion from toxic text to neutral text makes this task complex. The baseline methods, such as keyword deletion, show reasonable results but are insufficient and prone to mistakes. They can unintentionally censor non-toxic content or fail to capture proper toxic messages. Therefore, a more sophisticated approach that understands context, nuances and nature of language is required. This could be achieved with the use of pre-trained sequence-to-sequence models.

2 Data

The ParaDetox dataset (Logacheva et al., 2022) is the parallel corpus dataset specifically designed for the detoxification of English texts. The dataset contains toxic messages with their corresponding neutral counterparts. Overall, the size of the dataset is 19,744 pairs. For the purposes of training and evaluation, the dataset was divided into training, validation and test sets using a 70/10/20 percent splits. This resulted into 13,820 pairs for training, 1975 for validation and 3949 for test set.

Before being fed into the model, the data was tokenized using a tokenizer suitable for the chosen model, converting sentences into a series of tokens that serve as model's input. Both padding and truncation were done to ensure equal sizes of inputs. The maximum lengths of toxic or neutral sentences were no more than 150 characters, with the average length of toxic sentences of 54.8 characters and 46.3 characters for neutral. The Table 1 shows some examples of the dataset.

3 Methods

In this project, I address the challenge of text detoxification via fine-tuning of sequence-to-sequence model based on the Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al., 2019). I fine-tuned and assessed two BART models, namely 'facebook/bart-base' and 'facebook/bart-

Toxic Sentence	Neutral Sentence
delete the page and shut up	please delete the page
honestly catcalling really sounds stupid .	It's not wise to catcall frankly.
it would be stupid not to .	it wouldn't be intelligent not to.
the police dont have time for this crap .	The police have no time for this.
i want you all to myself , you sexy beast .	i want you all to myself , honey .

Table 1: Few examples from the dataset.

large'.

To evaluate the effectiveness of my detoxification model three different evaluation metrics were used. These evaluation metrics were proposed by dataset authors and can be accessed through cloning their github. The metrics are:

1. Style Transfer Accuracy (STA): This metrics measures the percentage of outputs that were successfully detoxified, as identified by a high performance toxicity classifier trained on Jigsaw dataset. So, this metrics shows how accurate the model is in removing toxicity of the text.

2. Content Preservation (SIM): Cosine similarity is used to measure how well the model preserves the original content after detoxification process. This involves computing the cosine similarity between the embeddings of the original toxic text and the detoxified output.

3. Fluency (FL): Fluency is measured by using perplexity by evaluating how likely a pre-trained language model considers a sequence of words to occur. This metrics is essential as it ensures that the transformed text remains grammatically sound and easy to understand.

Together these metrics provide a comprehensive framework for evaluating text detoxification model.

4 Experiments

4.1 Tools used and training regimen

The training procedure was completed using HuggingFace's Seq2SeqTrainer class. Tokenized toxic sentences were used as inputs to the model and tokenized neutral sentences were used as labels. DataCollator was used to pad the sequences to the maximum length.

In order to observe the performance during training, the evaluation metrics were applied on the validation set. However, due to the compatability problems with text embedding models provided by dataset authors and resource limitations, I used ROUGE metrics instead, where ground truth neutral sentences were compared to sentences generated by the model. But, the final evaluation on test set was performed using proper Content Preservation metrics.

4.2 Hyperparameters

For Seq2SeqTrainer, the following parameters were used:

- Batch size: 128
- Number of epochs: 10
- Optimizer: AdamW
- Learning rate: 5e-5
- Weight decay: 0.01

The training followed an epoch-wise evaluation strategy, assessing the model's performance on the validation set after each epoch.

In addition to the initial model, a fine-tuning process was also applied to the *bart-large* model using the same hyperparameters. This model, which is a larger variant of the standard BART model with more parameters and thus a higher capacity for learning complex patterns, demonstrated superior results on the validation set. The implication here is that the increased model size can more effectively capture nuances in the detoxification task, leading to better performance when generating neutral text from toxic input.

Adjustments to the learning rate were explored as part of the hyperparameter tuning process. It was observed that reducing the learning rate below the established 5e-5 led to significantly slower convergence rates. Specifically, the Style Transfer Accuracy (STA) was only at 60 percent in the initial epochs when a lower learning rate was used. Conversely, with a learning rate of 5e-5, the STA demonstrated high accuracies from the first epochs, indicating a more efficient and effective training

process.

Thus, the chosen learning rate of 5e-5 appears to be optimal for this particular model and dataset, offering a good compromise between speed of convergence and model performance.

5 Results

The evaluation of the text detoxification model was conducted using the following metrics on the test set:

Style Transfer Accuracy (STA): Style Transfer Accuracy was computed using Roberta classifier fine-tuned to detect toxic sentences. Higher values indicate better performance in achieving the desired non-toxic style.

Content Similarity (SIM): The sentence embeddings for contnent similarity were calculated by a model provided from dataset authors. High similarity scores indicate that the model is effective at maintaining the original message and meaning.

Fluency: Calculated using token-level perplexity with *GPT-2 medium* (Radford et al., 2019). Lower perplexity indicates more natural, fluent text that is likely to be similar to human-written language.

The training procedure for both models was steady in terms of reducing training loss and improving the performance on the validation set. The training could further be improved by extending the number of epochs and further hyperparameter tuning.

Two baseline methods were employed for comparative purposes:

Toxic Sentence Duplication: This method simply duplicates the toxic sentence without any transformation, serving as a control to demonstrate the necessity of detoxification.

Swear Word Removal: For this baseline, a *better-profanity* python library was used to identify and remove swear words from the sentences, providing a rudimentary method of detoxification.

The predictions generated by the fine-tuned *bart-base* and *bart-large* models were compared against the baseline methods. The Table 3 shows few examples of model outputs used in this study.

As shown in Table 2 the STA for the *bart-large* model indicated superior style transfer capability with Style Transfer accuracy of 90.3% compared to the *bart-base* model's 89.1%. This shows that the *bart-large* is more successful at neutralizing toxic sentences.

In terms of content similarity, the BART models showed approximately the same scores (88% for bart-base and 87% for bart-large). This shows that these models are effective in transforming toxic language to neutral style while maintaining the content's intent. The swear word removal method had a higher SIM score than both BART models, indicating that it maintained content similarity more effectively, as it only removes offensive words without altering the rest of the text. However, its STA was lower, showing that simply removing swear words does not fully neutralize the style of toxic sentences. Sentence duplication had an STA of 0.0, confirming its ineffectiveness in style transfer, as it does not alter the toxic content. It also shows the effectiveness of the toxicity classifier model used for STA metrics.

Even though, *bart-large* showed better detoxification performance than *bart-base*, it recorded a higher perplexity score, indicating less fluency. This unexpected result could suggest that the larger model is producing more complex or less common language constructs that are not as well recognized by the GPT-2 metric used for evaluating fluency. *bart-base*, while slightly less effective in style transfer, produced more fluent output according to perplexity score.

The sentence duplication baseline has a perfect SIM score of 100, as it makes no changed to the input text. Its fluency scores is the lowest, which suggests that the original sentences were generally fluent but toxic.

The swear word removal baseline shows the worst fluency score among all methods, which can be explained by the loss of important sentence structure due to the absence of necessary words. The fluency score could probably be better if these swear words were replaced with non-toxic synonyms.

The baseline methods, while serving as a necessary point of reference, demonstrated clear limitations. Toxic sentence duplication failed to address toxicity, and swear word removal, although partially effective, often resulted in sentences that were not grammatically correct or coherent.

While *bart-large* shows a slight advantage in style transfer accuracy over *bart-base*, it seems to do so at the cost of fluency. *bart-base*, therefore, offers a better compromise between detoxifying content and maintaining fluency, making it potentially more suited for applications where the naturalness of the text is a priority.

By manual observation of sentence structure of

bart-large model's outputs, it is difficult to observe the fluency problems that make up to a very high number of 819.69. It is hard to observe because visually these sentences seem to be fluent to some extent. If we compare the predictions by *bart-large* and *bart-base*, we can observe that these models produced 2,290 identical predictions out of 3,949 sentences in test set. Consequently, the fosuc shifted to examining the biggest differences in perplexity values between the outputs of both models. For instance, the sentence "the real world is complicated?" from *bart-base* yielded a perplexity of 210.5, whereas, *bart-large* generated "The real world is complicated." with a significantly lower perplexity of 26.0. This difference can be attributed to the questionable use of "?" sign in an otherwise identical sentence, which is not typically expected by the gpt-2 medium model. Despite the "?" being in the original sentence, this example demonstrates that perplexity evaluations can yield misleading results based on mere punctuation differences. Another example further illustrates this point: *bart-base* produced "What is going on?" with perplexity of 356.5, while *bart-large* generated "what is going on??!" with a perplexity of 622.8, showing how additional punctuation marks dramatically increase the measured perplexity.

In most of the cases, the perplexity is a reasonable measure of fluency, but it can also produce potentially misleading results when evaluating sentences where minor differences, such as punctuation usage, are present. This results in a problem of properly comparing perplexity results of different models.

6 Conclusion

This paper presented a comprehensive study on text detoxification using advanced sequence-to-sequence models, specifically the BART models. Through fine-tuning and evaluation, this study aimed to transform toxic sentences into neutral sentences while retaining the original content and ensuring fluency. The bart-base model, while slightly less effective in style transfer, provided outputs with greater fluency. This model represents a more balanced approach, potentially better suited for real-time applications where the naturalness and flow of conversation are critical.

Nonetheless, our results also point to the need for more refined fluency metrics. The current use of GPT-2 perplexity, although informative, may not capture the full spectrum of language nuances. Future work should focus on developing or identifying more sophisticated fluency metrics that can more accurately reflect human evaluations of text naturalness.

Moreover, it is possible that extending the training duration beyond the current number of epochs could lead to enhanced model performance. Further epochs would provide the models with more opportunities to learn from the training data, possibly improving both style transfer accuracy and fluency.

References

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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

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

Models	STA	SIM	Fluency
duplication	0.0	100.0	342.38
swear words removal	81.8	93.6	1116.57
bart-base	89.1	88.7	374.97
bart-large	90.3	87.3	819.69

Table 2: Test set evaluation results.

Original	anonymous is so fucking annoying.	where tha fuck dis rain kome frm	holy shit the real world is complicated ?
bart-base	anonymous is so annoying.	Where did this rain come from?	the real world is complicated?
bart-large	anonymous is so annoying.	where dis rain come from?	The real world is complicated.
removal	anonymous is so annoying .	where tha dis rain kome frm	holy the real world is complicated ?

Table 3: Conversion examples.