Name: Made Oka Resia Wedamerta
Email: m.wedamerta@innopolis university

Final Solution Report

1. Introduction

The detection of toxicity in user messages is a current research topic [1]. The process of automatic rewriting of offensive text drew less attention, but it might have a variety of beneficial uses, such as ensuring the internet is more peaceful by suggesting to a user that they submit a more neutral version of an emotive comment. Current research on text detoxification [2] characterizes this process as style transfer. The style transfer job is commonly defined as recreating text with the same content but changing one or more of the qualities that comprise the "style," such as authorship [3], emotion [4], or level of politeness [4]. Despite the purpose of retaining the text, modifying the style features often drastically affects the meaning of a statement. In reality, many style transfer models aim to change a text into a somewhat comparable sentence written in another style on the exact same subject. The final solution report presents the culmination of efforts in text generation using T5 models. The task at hand involves transforming toxic comments into neutral comments. This report provides a comprehensive overview of the model, training process, and evaluation criteria.

2. Data Analysis

The data analysis phase plays a crucial role in understanding the dataset's composition and characteristics. It encompasses the following key aspects: Data Collection, Data Distribution, and Exploratory Data Analysis (EDA) which is conducted to reveal insights into the nature of toxic and neutral comments, including word frequencies, lengths, and patterns. First of all, I use The dataset is a subset of the ParaNMT corpus and the dataset comprises pairs of toxic and neutral text which is translation of the toxic text. Furthermore, The distribution of toxic and neutral comments is analyzed to identify any class imbalances. I found that there is some problem in the dataset. Among 500K sentence pairs, more than half of the data have a strange distribution. To be more precise, the translation text has a higher toxicity score than the reference (toxic text). Furthermore, after assessing it manually, I recognize that the translation text and the reference text seem like swapped, meaning that the reference text is supposed to be in the translation and vice versa. Finally, as data preprocessing I swap the reference text and score with the translation text and score if the specific data meet the condition reference toxicity score is smaller than translation toxicity score.

3. Model Specification

The model used for this task is the T5 (Text-To-Text Transfer Transformer) architecture, specifically 't5-small.' The T5 model is known for its capability to perform various NLP tasks, and it is fine-tuned for the specific task of generating neutral comments from toxic comments. The model's include AutoTokenizer from the Hugging Face Transformers library as tokenizer

4. Training Process

The training process involves several key steps. First step the dataset is split into training and testing subsets to evaluate the model's performance effectively. Then I use PyTorch's DataLoader to efficiently load and manage data batches for model training. For handling padding and collation of input features I implement the DataCollatorWithPadding class. The second step is the training loop, in which the model undergoes training with specific hyperparameters, including batch size, maximum epochs, and maximum steps. The training loop includes gradient accumulation for enhanced stability. The model is optimized using the Adam optimizer and the Exponential Moving Average loss is employed to monitor and record the model's loss during training.Finally, to allow resuming training and model deployment, the model checkpoints are saved periodically.

5. Evaluation

Because there's no parallel set of tests to perform the detoxifying job, we cannot utilize BLEU, METEOR, or ROUGE metrics and must rely on reference-less assessment [7]. Style transfer models must modify the style while preserving the information and producing a fluid text. Because these factors are frequently inversely connected, we require a compound metric to establish an equilibrium among them. We utilize the measure J, which is the multiplication of sentence-level style accuracy, content preservation, and fluency, and we follow the assessment technique of [6]. The average of sentence-level scores is used to get the system-level J. A pre-trained toxicity classifier is used to assess style accuracy (ACC). The cosine similarity of sentence-level content preservation (SIM) is measured. The average of their sentence-level product is used to calculate J.

6. Result

The model presented in this solution demonstrates its capability to generate less toxic sentences from the given toxic inputs. The model is successful in partially detoxifying the provided toxic sentences. For instance, a sentence like "you are an idiot" may be transformed into a less toxic version such as "you are a fool." This showcases the model's potential to make the language more polite and less offensive. One potential issue affecting the model's performance is the number of training steps. The model's training was limited to 10,000 steps. While this allowed for a basic evaluation of the model, it is possible that the model's performance could improve with a more extended training period. Setting the step count to around 26,000 might enable the model to reach higher levels of proficiency. Although the model achieves some level of detoxification, it's important to acknowledge that it may not fully detoxify all toxic sentences. Achieving complete detoxification remains a challenging task, and further model development is required to approach the performance of state-of-the-art (SOTA) models, as demonstrated in [7]. In conclusion, this solution offers a valuable step forward in addressing toxic content by generating less harmful sentences. However, there is room for improvement, particularly in terms of training steps. Future iterations of the model may aim for more extended training periods to further enhance detoxification capabilities and approach the level of SOTA models in the field.

7. Reference

[1] M. Zampieri *et. al*, "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)", *In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1425– 1447, Barcelona (online)*, 2020, International Committee for Computational Linguistics.

[2] L. Laugier, J. Pavlopoulos, J. Sorensen, and L. Dixon, "Civil rephrases of toxic texts with self-supervised transformers", 2021, CoRR, abs/2102.05456.

[3] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov, "Rtgender: A corpus for studying differential responses to gender", *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA)*.

[4] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment", *In Advances in Neural Information Processing Systems, volume 30*, *2017, Curran Associates, Inc*.

[5] A. Madaan *et. al*., "Politeness transfer: A tag and generate approach", *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *pages 1869–1881, Online. Association for Computational Linguistics*, 2020.

[6] K. Krishna, J. Wieting, and M. Iyyer, "Reformulating unsupervised style transfer as paraphrase generation", *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 737–762, Online*. 2020 Association for Computational Linguistics.

[7] D. Dale *et. al*., "Text Detoxification using Large Pre-trained Neural Models." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, doi: 10.18653/v1/2021.emnlp-main.629.