Text Detoxification

Introduction

Text detoxification is a task in natural language processing (NLP), where the goal is to transform toxic or harmful text into non-toxic content while preserving the original context and meaning. In this report, we present our approach to text detoxification, which leverages two distinct machine learning models to achieve this goal. Our method involves sampling from a paraphrasing BART (Bidirectional and Auto-Regressive .Transformers) model and selecting the least toxic result using a BERT-based classifier

Problem Statement

The challenge of text detoxification arises in various contexts, such as content moderation on social media platforms, ensuring the safety of online communication, and improving the overall quality of text data. Toxic or harmful text can include hate speech, offensive language, and harmful content, which pose risks to users and platforms alike. Our method aims to mitigate these risks by automatically detoxifying .text

Methodology

Our approach to text detoxification involves two key components: a paraphrasing BART .model and a toxicity classifier based on BERT

Paraphrasing BART Model

We begin by utilizing a BART model fine-tuned for paraphrasing tasks. This model is pre-trained on a large corpus of text and then fine-tuned on our dataset for text detoxification. The core idea is to leverage the paraphrasing capabilities of this model to generate alternative, non-toxic versions of the input text. We fine-tuned the model on . our dataset so that it learns to generate non toxic paraphrasing

Sampling from the Paraphrasing BART Model

Input Text: We start with a given input text that may contain toxic elements

Paraphrase Sampling: We sample multiple paraphrases of the input text from the BART paraphrasing model. This step allows us to generate a variety of alternative .versions of the original text, each with potentially different phrasing and wording

Generation Process: The paraphrasing BART model generates these alternative versions based on its pre-trained knowledge of language and context

Output Candidates: After paraphrasing, we obtain a set of candidate non-toxic texts. .However, not all candidates may be equally effective in detoxifying the text

Toxicity Classifier BERT

To identify the least toxic result from the generated candidates, we employ a BERT-based toxicity classifier. This classifier is specifically trained to distinguish between toxic and non-toxic text. By using this classifier, we can rank the generated candidates and select the one with the lowest predicted toxicity score. this ensures that our model .doesn't generate toxic text while preserving the meaning

Results

Our method for text detoxification using dual models has shown promising results, effectively transforming toxic text into non-toxic content while preserving the original context. Here are some key findings from our experiments on BART model without fine :tuning showed

Accuracy: Our approach achieved an accuracy of 87% in detecting toxic text, indicating that it effectively identified and selected non-toxic candidates.

Preservation of Context: By leveraging the paraphrasing BART model, we were able to retain the context and meaning of the original text, ensuring that the detoxified version .remains coherent

Efficiency: The use of a pre-trained paraphrasing BART model allowed us to efficiently generate multiple candidate versions, and the BERT-based toxicity classifier efficiently .ranked and selected the least toxic candidate

Conclusion

Text detoxification is a critical task in ensuring safe and respectful communication in various online platforms and applications. Our method, which combines the power of a paraphrasing BART model with a toxicity classifier based on BERT, demonstrates the potential for effective detoxification while preserving context and meaning. Future work may involve further optimizations, including model fine-tuning, dataset expansion, and ..addressing potential challenges in handling complex language scenarios