ANLP Interm Report

Team FatGPT

Atharva Joshi
2020111010

Shavak Kansal
2020101023

Dhruv Mittal
2020113017

I. Problem Statement

In the age of digital communication, the proliferation of harmful and toxic language within text-based content has emerged as a formidable societal issue. Toxic text, comprising hate speech, offensive remarks, and harmful content, is pervasive across online platforms, leading to a toxic digital environment that poses real threats to individual well-being and community cohesion. Addressing this problem necessitates the development of robust automated text detoxification methods that can process and rehabilitate toxic text to make online spaces safer and more inclusive.

The primary objective of this project is to tackle the problem of unsupervised text detoxification, wherein toxic text is identified and systematically transformed into non-toxic, socially acceptable language, all while retaining the underlying meaning and context. To achieve this, we will draw inspiration from the CondBERT approach, a novel methodology that capitalizes on transformer encoder models for the purpose of generative token replacement. This token-by-token replacement approach is designed to ensure that the detoxified text remains coherent and semantically accurate.

The project will confront several pivotal challenges:

1. **Toxic Text Identification:** Accurately recognizing toxic tokens within text is a fundamental challenge. We must develop mechanisms to identify and classify toxic segments efficiently.

2. **Token-by-Token Replacement:** Once toxicity is identified, the model must replace toxic tokens with non-toxic alternatives. This process necessitates a deep understanding of the context to avoid unintended distortions.

3. **Unsupervised Learning:** We will employ an unsupervised learning approach, reducing the reliance on labeled data and instead exploiting the knowledge encoded in pretrained transformer models.

4. **Meaning Preservation:** Preserving the original meaning and intent of the text is paramount. The model must guarantee that detoxified text does not misrepresent the author's message.

5. **Evaluation and Baseline Comparison:** The project will develop robust evaluation metrics to measure the effectiveness of the CondBERT-based approach. Comparisons against existing pointwise editing techniques for style transfer will be conducted.

II. Progress

A. Toxic Sentence Classification with RoBERTa

To identify and mask toxic words in a sentence, we have employed a toxic sentence classifier based on RoBERTa. The approach we have implemented is outlined as follows:

- For a toxic sentence consisting of n words, we create n variations of the sentence by masking a different word in each variant.
- We then run these variants through the classifier to identify the sentence with the least toxicity.

The code snippet below illustrates this process:

```
batch = tokenizer.encode(sentence, return_tensors='pt')
output = torch.nn.functional.softmax(model(batch).logits
toxic_score = float(output[0][1])
if min_tox > toxic_score:
  min_tox = toxic_score
  masked_sentence = sentence
```

We have established a self-defined threshold, which in this case is set at 0.25. We check if the masked sentence's toxicity score is now below this threshold. If not, we iterate through the process again, masking one more word in each iteration, until we obtain a masked sentence with a toxicity score less than the specified threshold value.

B. Style Transfer using baseline models

We Planned on evaluating our dataset against the baseline models: T5 Paraphraser and DRG-RetrieveOnly.

T5 Paraphraser: In the paper Reformulating Unsupervised Style Transfer as Paraphrase Generation, it was found that paraphraser was also able to reduce the toxicity of a sentence to a certain extent. This paraphraser was fine-tuned on model PAWS (Paraphrase Adversaries from Word Scrambling) , MSRP(Microsoft Research Paraphrase Corpus), and Opinosis.

DRG-RetrieveOnly: A straightforward yet effective approach to style transfer involves keeping the sentence unchanged and only modifying individual words that are linked to the desired style. DRG-RetrieveOnly retrieves a sentence with the opposite style which is similar to the original sentence and returns it.

However, due to lack of time and very high training time, we were not able to obtain the results for DRG-RetrieveOnly

III. Baseline Results

T5 Paraphraser: The metric used to evaluate the model was the mean product of similarity, fluency and perplexity. The score obtained was **4.370629371e-3**

DRG-RetrieveOnly: The code for evaluating this model has been written but due to very long running times, the model could not be trained completely and results could not be obtained.

References

[1] David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text Detoxification using Large Pre-trained Neural Models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[2] Isuru Gunasekara and Isar Nejadgholi. 2018. A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 21–25, Brussels, Belgium. Association for Computational Linguistics.

[3] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67.