Report on Solution Building for Text Detoxification

Artyom Makarov

Introduction

Text detoxification, the process of converting toxic or offensive text into a non-toxic form, is a challenging problem in natural language processing. In this report, I suggest three hypotheses for addressing this problem and evaluate their effectiveness. I also provide a comparison of the resulting models based on semantic similarity, style accuracy, fluency, and J score, introduced in [2] (Krishna, Wieting, & Iyyer 2020)

Hypothesis 1: BERT-Based Masked Language Model

As the first hypothesis I suggested using a Masked Language Model based on BERT to replace toxic words with [MASK] tokens. While this approach is straightforward and does provide some level of detoxification, it has limitations. It lacks the context to identify toxic phrases, and replacing words with [MASK] can lead to the loss of sentence meaning. Nevertheless, it serves as a reasonable starting point for this problem.

Hypothesis 2: Toxicity Classifier with BERT Masked Language Model

Hypothesis 2 proposes training a toxicity classifier to identify toxic words and then replacing them with [MASK] tokens. This approach is theoretically more robust, but in practice, it performs poorly. The classifier's training data is unbalanced, leading to the misclassification of common words as toxic, resulting in entire sentences being replaced with [MASK] tokens. This approach also suffers from the same limitations as the first hypothesis. That has encouraged me to take a step in a different approach.

Hypothesis 3: T5-Small Model for Style Transfer

The third hypothesis introduces a different approach by fine-tuning the T5-Small Model for style transfer. This method eliminates the need to replace words with [MASK] tokens or train a toxicity classifier. Instead, it feeds the training data directly to the model, which then generates non-toxic equivalents. Although not perfect, this approach demonstrates impressive results, as it understands the context of toxic words and replaces them with non-toxic alternatives.

Some examples of detoxified text:

Original: *I can't believe we haven't fucked for two years, nine months, three weeks and... 69 hours.*
Detoxified by Hypothesis 1: *i can't've we haven't been ed for two years, nine months, three weeks and... 69 hours.*
Detoxified by Hypothesis 2: *i cannot believe i just fucked up two years, nine months, two weeks and... 69 hours.*
Detoxified by Hypothesis 3: *I can't believe we haven't slept for two years, nine months, three weeks and... 69 hours.*

Original: *So forgive me for being a little fidgety, but if it's lunch we're talking, I'm gonna eat a fat pork sandwich, and I'm sure as shit not gonna eat it here.*
Detoxified by Hypothesis 1: *so forgive me for being a little co y, but if it's lunch we're talking, i'm gonna eat a roast pork sandwich, and i'm sure as well not gonna eat it here.*
Detoxified by Hypothesis 2: *so forgive me to just a little. but if your lunch stop talking, i'm to eat a c ing sandwich, and i'm good as good as you eat right here.*
Detoxified by Hypothesis 3: *so forgive me for being a little bit of a snitch, but if it's lunch we're talking, I'll eat a fat pork sandwich, and I'm sure I'm not gonna eat it here.*

Original: *There is no fucking soy milk!*
Detoxified by Hypothesis 1: *there is no - ing soy milk!*
Detoxified by Hypothesis 2: *there is no real soy.*
Detoxified by Hypothesis 3: *there's no soy milk!*

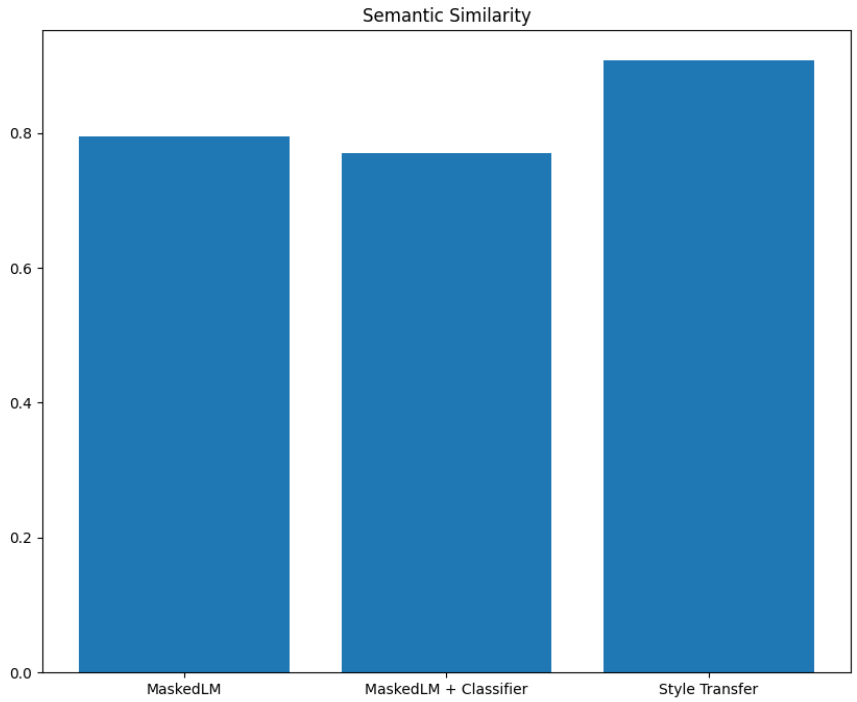
Original: *What's wrong with people having sex?*
Detoxified by Hypothesis 1: *what's it with people having babies?*
Detoxified by Hypothesis 2: *what's the about people had sex?*
Detoxified by Hypothesis 3: *what's wrong with people having sex?*

Original: *What the fuck are you talking about?*
Detoxified by Hypothesis 1: *what the world are you talking about?*
Detoxified by Hypothesis 2: *what the world are you talking about?*
Detoxified by Hypothesis 3: *what are you talking about?*

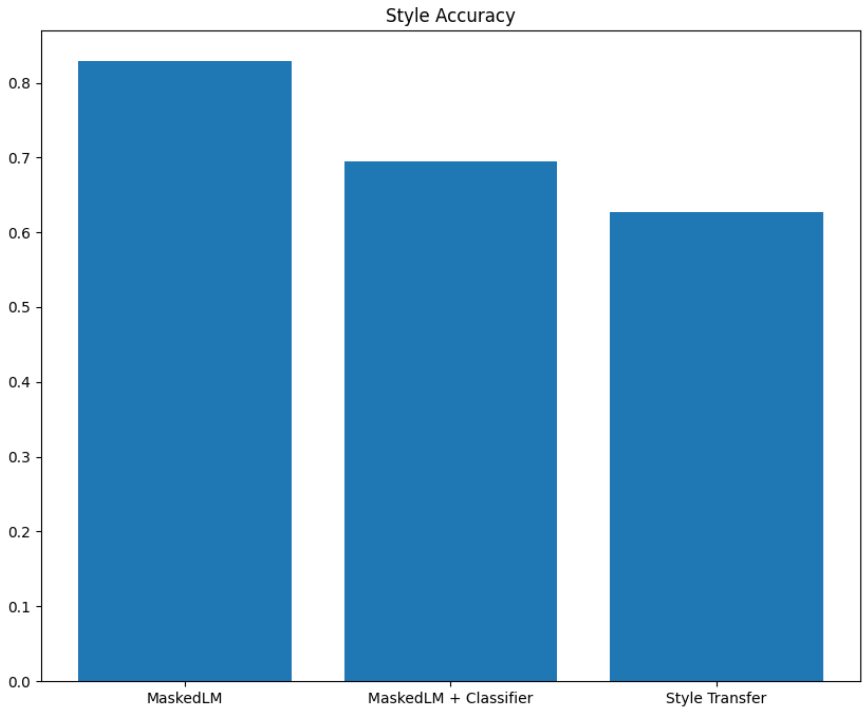
Results and Model Comparison

To assess the effectiveness of the three hypotheses, I evaluated the resulting models using four key metrics: semantic similarity, style accuracy, fluency, and a J score.

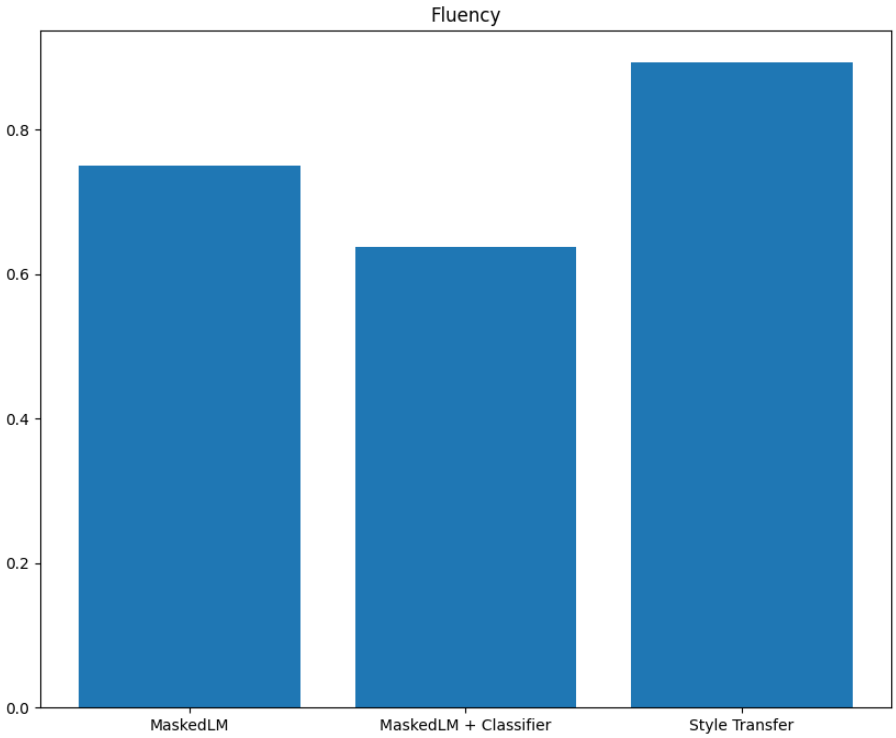
- **Semantic Similarity**: This metric measures how closely the detoxified text resembles the original text in terms of meaning. It can be seen that Hypothesis 3 outperforms the other two hypotheses.



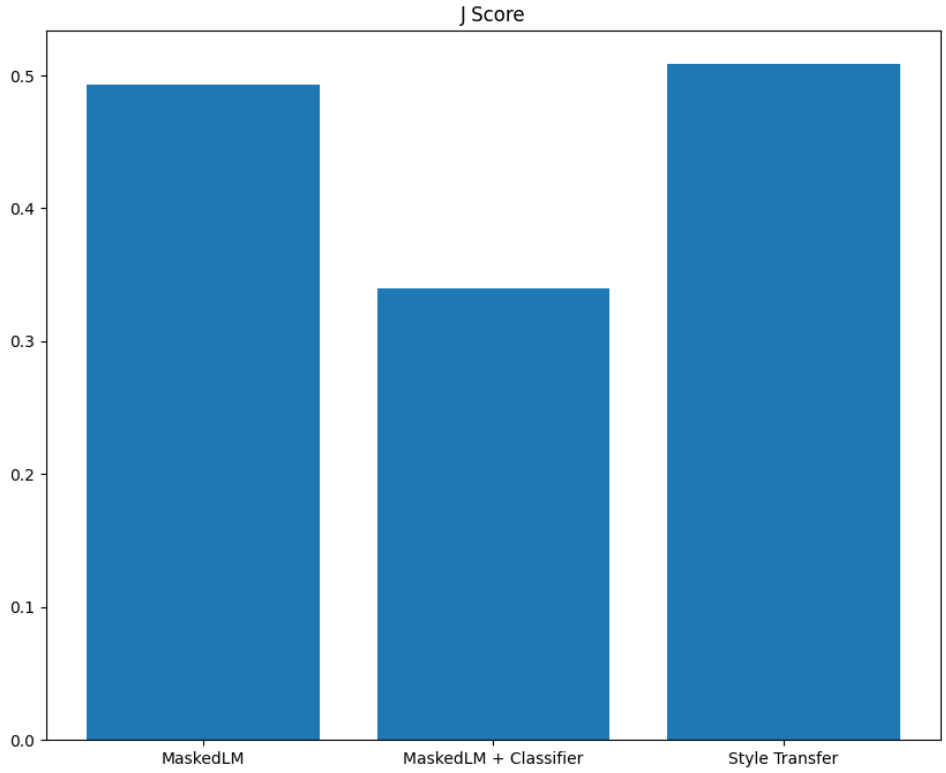
- **Style Accuracy**: Style accuracy evaluates how well the text was detoxified. Hypothesis 1 scores highest in style accuracy, as it completely replaces toxic words with [MASK] tokens.



- Fluency: Fluency assesses how well the detoxified text is structured according to the English grammar. Hypothesis 3 excels in fluency due to the T5-Small Model's ability to generate coherent sentences.



- J Score: The J score combines semantic similarity, style accuracy and fluency. Hypothesis 1 and 3 had J scores that were very close, with both outperforming Hypothesis 2.



Conclusion

In conclusion, while it is worth noting that the Hypothesis 1 model performs nearly as well as the Hypothesis 3 model according to the J metric, it is important to emphasize that the sentences generated by the third model have a more "human" quality. Hypothesis 1 may excel in certain metrics like style accuracy, but Hypothesis 3's outputs show a better understanding of context and coherence, contributing to a more natural and human-like text transformation.