SmartCensor: A Transformer-based Detoxifier for Toxic Sentences

David Simonetti, John Lee

Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, US.

Motivation and Goal

Since the inception of the internet, toxic comments have proliferated on the web. Chat rooms, social media, and online video games have always been flooded with this hateful and malicious content. In order to make the internet a less toxic place, we wish to create a NLP model which is able combat online toxicity. We envision the model taking in arbitrary chats from users, which may contain obscenities, racism, sexism, etc, and outputting a sanitized text which will communicate the same idea without all the hate. The idea is that other companies can plug this model into their online communication systems. For example, user chats in a video game match would pass through this model before being sent to other players. This would make interacting with others online a much less negative experience than before.

Datasets

We used three datasets in the training and evaluation of our model

Paradetox – Parallel Detoxification Data
This is a dataset of 19,700 entries consisting of an original toxic sentence and a detoxified version of the same sentence intended to have the same meaning. We used this dataset to train our model as an instance of machine translation, i.e. our model was trained to translate a toxic sentence into a detoxified sentence. We also used 20% of this dataset as a dev set to validate the model during training.

Example:
Obama has been a total failure , and now looks like a sore loser. →
Obama has not been victorious.

Jigsaw - Wikipedia Toxic Comment Dataset
This dataset consists of comments taken from Wikipedia that have been labeled by human moderators as one of the following: toxic, severely toxic, obscene, threatening, insulting, or hateful of identity. We use roughly 16k test samples from this dataset to evaluate both the baseline and our trained model.

Example:
Chris, you mother fucker...all what you want to know about ChrisO you can find at www.ChrisO.homo.com -> toxic, obscene, insult

Toxic Word Bank - Dictionary of Toxic Words
This dataset is a curated list of known toxic words. It contains pure curse words, slurs, as well as common slangs and abbreviations of words used to harm others. We use this in our baseline in order to censor the toxic content of the sentence by searching for and removing all words in this word bank.

Experimental Results

Metric	Original Averages		Baseline - Replacement		SmartCensor	
	All	Toxic Only	All	Toxic Only	All	Toxic Only
Toxicity	0.1022	0.8661	-0.0250	-0.3696	-0.0615	-0.6895
Severe Toxicity	0.0152	0.1437	-0.0111	-0.1061	-0.0148	-0.1411
Obscenity	0.0644	0.5926	-0.0400	-0.3857	-0.0613	-0.5775
Threat	0.0065	0.0598	-0.0032	-0.0302	-0.0046	-0.0439
Insult	0.0614	0.5705	-0.0344	-0.3400	-0.0553	-0.5334
Identity Attack	0.0128	0.1183	-0.0072	-0.0719	-0.0112	-0.1054
Cosine Similarity	N/A	N/A	0.7103	0.7096	0.3733	0.3928

Figure 1: Comparisons in toxicity and cosine similarity on Jigsaw Dataset between baseline and SmartCensor. Toxicity values are reductions.

Figure 2: Perplexity over Training

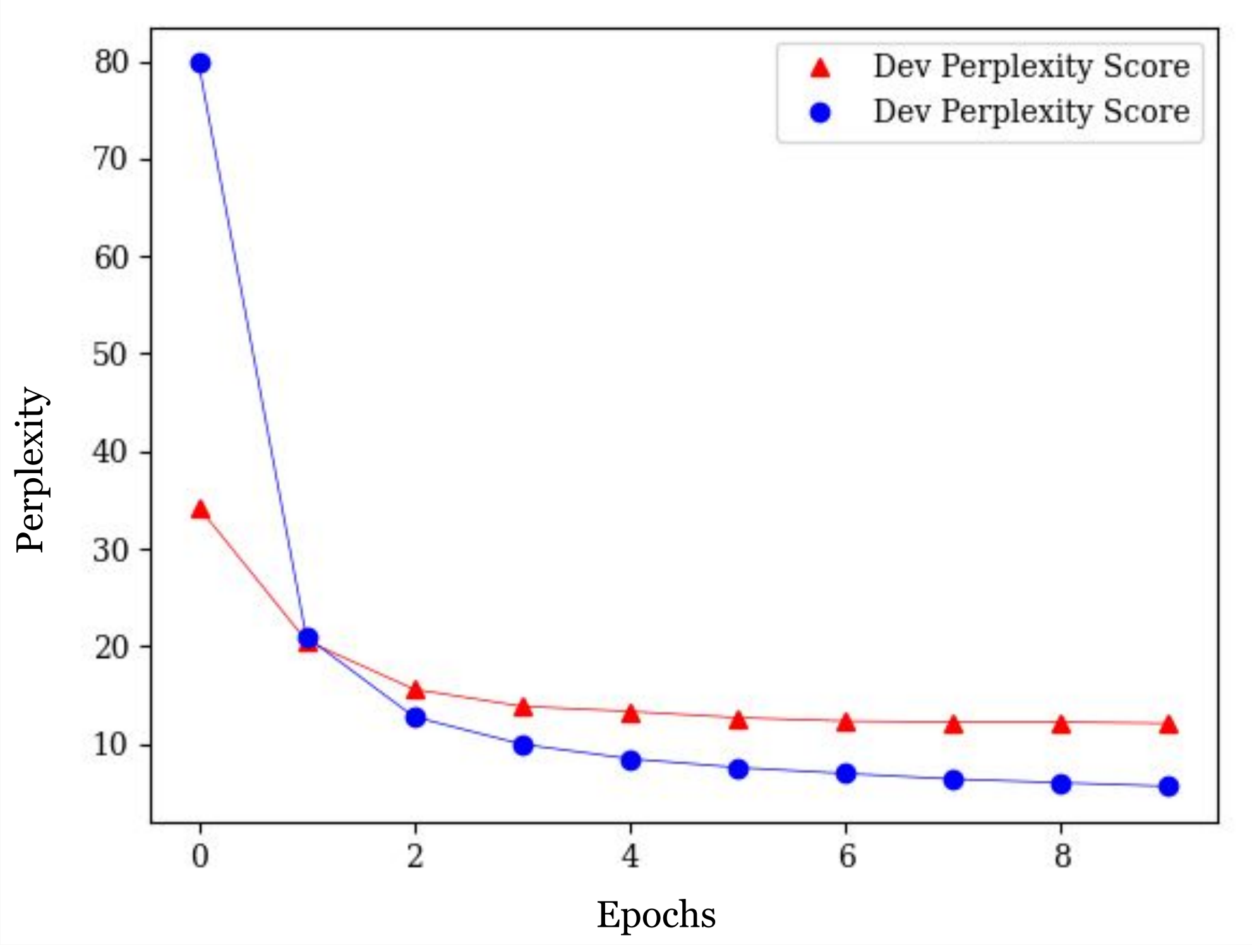
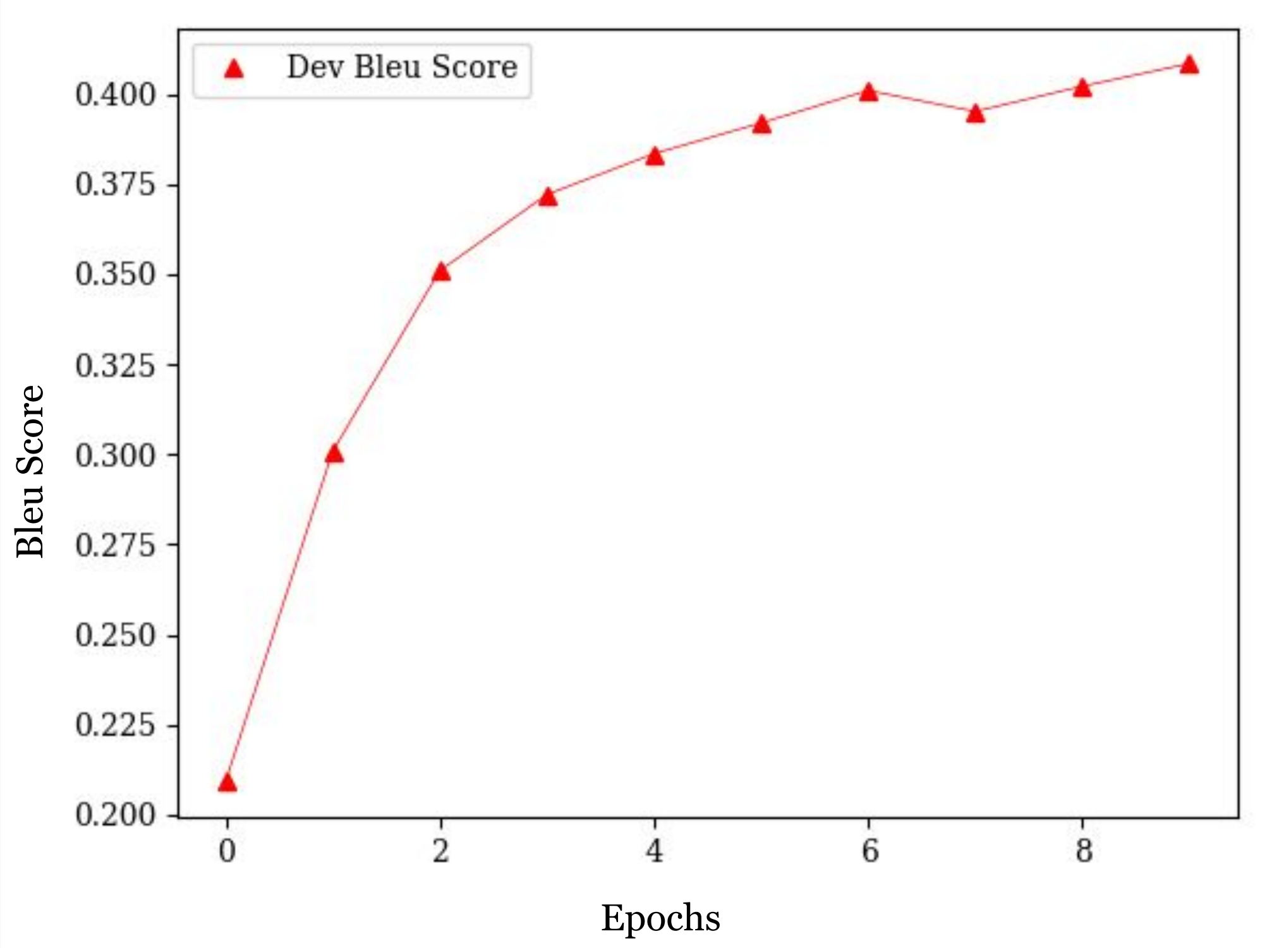


Figure 3: Dev Bleu Score over Training



Original	Detoxified	Truth
They treat you like sh*t.	They treat you badly.	They mistreat you.
still really f*cked up regardless .	still really messed up regardless of	still really messed up regardless
no , they haven't done jack sh*t with regard to isis .	No, they haven't done anything with jail.	No they haven't done anything with regard to isis.
f*ck now that song is stuck in my head.	I that song is stuck in my head	damn this song is stuck in my head.
urban warfare is a b*tch.	It is a bad person	urban warfare is bad

Figure 4: ParaDetox dev sentence detoxification

Original	Baseline	SmartCensor
f*ck you cr*cker nerd	you er nerd	I don't like you
STFU. I don't like you. Go away.	STFU. I don't like you. Go away.	It's It do wrong like you. It is.
This is your last warning. Next time, I will force you to eat your own sh*t.	This is your last warning. Next time, I will force you to eat your own.	It is your last warning It will force you to eat your own stuff.
MARK SWEEP IS A W*NKER	MARK SWEEP IS A ER	It is not good

Figure 5: Jigsaw test sentence detoxification

Conclusions

1. Despite the training data containing less rigorous examples of toxicity, SmartCensor performs better at removing toxicity than just a straight filter.
2. However, the lack of vocabulary recognition creates somewhat poor or boring translations. They are visually more coherent than the baseline results, but no good metric of coherency is provided.
3. The lack of vocabulary becomes particularly critical when forcing translations of non-toxic sentences.

Methodology

Encoder – We use a Transformer architecture for encoding. Words are first converted to word and positional embeddings (max length = 64) with a dimension of 256. The encoding layers consist of 4 alternating Multi-Head Self Attention Layers and residually connected Feed Forward Networks (FFN). The MH Attention layers have 4 heads each, and transform 256 dimensions into 256 dimensions. The FFNs have 256 input and output neurons, and 1024 hidden neurons.

Decoder – The decoder also creates word and position embeddings of input sentences like the encoder. These embeddings are fed through a Masked Self Attention Layer with 256 dimensions and a FFN with the same parameters as in the Encoder. The output of these embeddings in combination with the Encoder output is fed through a Cross Attention Layer of 256 dimensions, and then to a Softmax layer to retrieve "probability" scores.

Training – Our training method was based on machine translation, using the Paradetox dataset. We trained our model for 10 epochs, using an Adam optimizer with a learning rate of 0.0003. The loss is the negative log probability that the generated sentence is the translation of the input sentence. The model was saved at the point the dev Bleu score was the highest.

Evaluation Metrics - Toxicity measurements were found using the Detoxify library, containing a pretrained toxicity identification model Toxic Bert that classifies levels of from 0 to 1. Cosine similarities were found by comparing original and detoxified sentences using the all-MiniLM-L6-v2 semantic search model created bv the Sentence Transformers team.

Limitations

Limited parallel dataset - There exists a very limited set of toxic data with detoxified transitions. Beside manual data gathering, a suggestion is proposed in Future Work.

Less words in vocabulary - When we train, we also lack exhaustive vocabulary, resulting in some odd translations. This could be addressed by training on non-toxic translations of X to X, but may worsen toxicity masking.

Limitations in toxicity identification - Paradetox contains much more straightforward toxic words, which may not capture all toxicity.

Limitation in evaluation - There is no well-defined metric for sentence fluency or preservation. Cosine similarity is a very limited approximation.

Future Work

Toxifier - It is possible to train a toxifier by swapping inputs and outputs.

"Residual" Translation - We could possibly preserve words directly from the input sentence as long as they are not toxic.

Pseudo-GAN architecture - We propose feeding the Toxifier with a non-toxic sentence, updating the toxifier, then feeding the toxic sentence to the DeToxifier to retrieve the original sentence. This could be a potential solution for overcoming the lack of parallel data.