Neutralizing Toxic Language in Online Communications Using Transformers

April 20, 2023

Mai La

Jesse He

Shivangi Pandey

Berkeley
UNIVERSITY OF CALIFORNIA

Disclaimer - Viewer Discretion is Advised

- Due to nature of this research, we have text which are toxic in nature and have profanity and offensive language
- We have masked those words in the presentation, but negative sentences are presented at its natural form

Overview

Research has shown that abusive, toxic content on social media is detrimental to mental health

Content moderation has its limits. Executed incorrectly looks like censorship

Objectives: Given a toxic sentence, detoxify that sentence without losing its meaning

- Preserve overall semantics of the content
- Convert to inoffensive style

Evaluation Metrics:

- ROUGE Scores
- NonToxicScore

Dataset

- APPDIA - Inoffensive Social Media Conversations Transfer Dataset
  - Parallel corpus
  - 2k offensive Reddit short comments and non offensive transferred style

- Jigsaw Toxic Classification Dataset
  - Collection of comments from Wikipedia, Reddit and social media
  - 16K of comments classified in different toxic categories, 143K of neutral comments
  - Classification Model: toxic (label=0) or non-toxic (label=1)
  - Extract NonToxicScore: Probability of the text being non-toxic

Models

Generative Language Model

--

- GPT2 Few Shot (Baseline)
- GPT2 Fine Tuned
- GPT-Neo 2.7B Few Shot
- **GPT-Neo 1.5B Fine Tuned**

Hypothesis: LM could generate non-toxic text

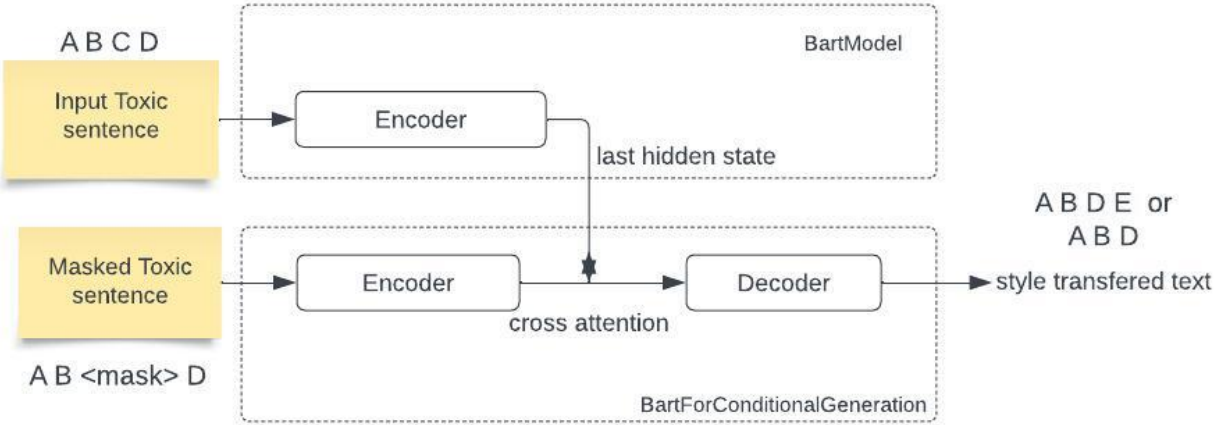
Encoder-Decoder Model

--

- T5 Fine Tuned
- BART Base
- **BART Large**
- BART Dual-Encoders

Hypothesis: analogous to a translation task

BART Dual-Encoders




```
NTA.    Dump his ****. Tablemanners are no rocket science. Treating other people like human beings is no rocket science. His still a child.
NTA.    Dump his <mask>. Tablemanners are no rocket science. Treating other people like human beings is no rocket science. His still a child.
```

Results - Classification

Evaluation on Jigsaw dataset:

	Max Length = 64		*Max Length = 512*	
Model	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**
BERT - 1 layer + Undersampling	90.8	93.1	-	-
DistilBERT - 1 layer + Undersampling	92.6	92.6	94.3	94.3
DistilBERT - 2 layers + Undersampling	93.1	93.0	**94.6**	**94.6**
DistilBERT - 2 layers + Upsampling + Augmentation	92.9	92.9	**94.4**	**94.4**

Results - Classification

Evaluation on APPDIA dataset:

Model: DistilBERT - 2 layers		*APPDIA Toxic Texts*		*APPDIA Non-Toxic Texts*	
Sampling Methodology	**Max Length**	**Accuracy**	**NonToxic Score**	**Accuracy**	**NonToxic Score**
Undersampling	64	**93.5**	9.2	66.3	64.5
Undersampling	512	91.0	9.1	63.3	63.0
Upsampling + Augmentation	64	89.0	**7.8**	82.4	78.4
Upsampling + Augmentation	512	88.9	8.4	**83.4**	**79.4**

Results

Evaluated Annotated Text Against Toxic Text				NonToxicScore on Annotated Text
Model	**ROUGE1**	**ROUGE2**	**ROUGEL**	
None (Non-toxic Annotated Text against Original Toxic Text)	68.9	56.1	68.3	79.4
Evaluated Generated Text Against Annotated Text				**NonToxicScore on Generated Text**
Model	**ROUGE1**	**ROUGE2**	**ROUGEL**	
Baseline - GPT2 - Few Shot	11.8	1.7	9.9	57.3
GPT2 - Fine tuned	27.8	21.1	27.5	67.4
GPT-Neo 2.7B - Few Shot	29.8	9.6	27.1	37.1
GPT-Neo 1.3B - Fine tuned	61.8	48.2	61.4	**74.9**
T5 - Fine tuned	66.8	54.7	66.4	39.3
BART base - Dual Encoders with Masking	59.8	47.9	59.5	57.3
BART base - Fine tuned	69.7	57.8	69.2	57.6
BART large - Fine tuned	**71.0**	**59.3**	**70.5**	58.2

Example

Toxic Text

I hope the bast**d suffered

Human

I hope the **guy** suffered

GPT-Neo Fine tune

I hope the **guy** suffered

GPT-Neo Few shot

I don't know if he suffered, but he's dead now.

GPT2 Few shot

I hope he didn’t suffer because he was a f***ing moron

GPT2 Fine tune

I’ve seen this before

BART Large

I hope the **guy** suffered

BART Base

I hope the bast**d suffered

BART Dual Encoders

I hope the pain you suffer has not been suffered

T5 Fine tune

I hope the bast**d suffered

Example 2

Toxic Text

the f***.... you are on drugs

Human

uh..... you are not being realistic

GPT-Neo Fine tune

What.... are you on drugs

GPT-Neo Few shot

the f***ing f*** are you on?

GPT2 Few shot

I don't care what you think about me, I'm here to help you out. I am here for you.

GPT2 Fine tune

ive been on drugs for years

BART Large

you are on drugs

BART Base

you are on drugs

BART Dual Encoders

the U.S. you are on drugs

T5 Fine tune

you are on drugs

Trade Off - Detoxification vs Semantic Preservation

Toxic Text

Have you ever had salt water? It's f***ing gross.

Human

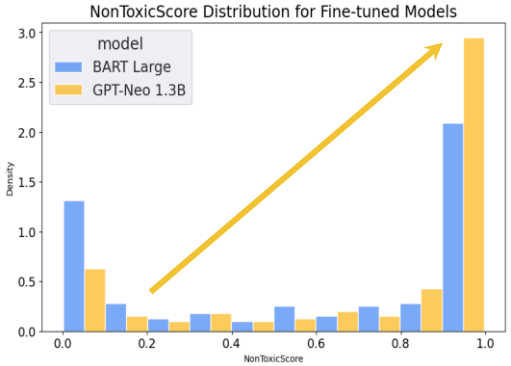
Have you ever had salt water? It's gross.

GPT-Neo Fine tune

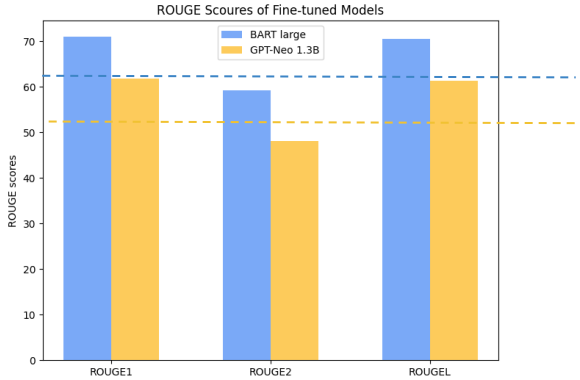
Have you ever had salt water? It's really bad.

BART Large

Have you ever had salt water? It's gross.



GPT: Better Detoxification



BART: Better Semantic Preservation

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusion & Next Steps

- Decoder only generative models produced more fluent, but sometimes semantically altered results
- Encoder-decoder models maintained semantic correctness, lower detoxification quality
- The optimal approaches involve tradeoffs, we aspire to eventually discover a solution that eliminates the need to compromise between semantic preservation and detoxification
- Next step: implement text-style disentanglement architecture

Thank you for your Attention()!

Appendix

NonToxicScore Calculation

Example Text	NonToxic Score	Predicted Label
Ok, this makes no sense. This will create vigilantes. And I hope they put this savage animal down before he hurts anyone else.	10.2	Toxic
This is really disgusting. People not able to compete even in the 3rd men league getting paid twice as much.	4.1	Toxic
Good lord you people are f***ing stupid.	0.1	Toxic
Yeah, no. And for that go away.	98.5	Non-Toxic
Holy moly this sub is gross	18.1	Toxic
Oh joy can't wait for all of you irrational people to misunderstand this science too. 🙄	51.3	Non-Toxic
Average NonToxicSCore	30.4	

Wrong Classification

Texts	Label	Predicted	NonToxic Score
There's plenty of evidence, it's exactly why they won't allow forensic audits *wipes voting machines clean of their records*	Toxic	Non-Toxic	97.6
Also there are women like this but they don't write about it like this on ducking reddit.	Toxic	Non-Toxic	99.0
Do you have a substantial counter argument or are you just spewing logical fallacies?	Toxic	Non-Toxic	87.3
Stop dude. Worst of the worst!	Non-Toxic	Toxic	33.3
They literally steal and sell babies in convenience stores. Sorry but most are undesirables. (a small % escape their culture and integrate)	Non-Toxic	Toxic	28.6
I would not say nuts...it just looks very suspicious. Either way, the dude is totally unacceptable.	Non-Toxic	Toxic	19.9

Sample Data

Sample prompt for few shot learning:

Rewrite the toxic text in non-toxic style:
###
Toxic text: So maybe you should be more retarded.
Non-toxic text: So maybe you should be more backward
###
Toxic text: You don’t deserve it dumbass.
Non-toxic text: You don't deserve it.
###
Toxic text: Youre soft as baby sh**
Non-toxic text: