Mitigating Bias and Toxicity

Arnab Bhattacharjee, Anirudh Ajith
Outline

1. What is toxicity?

2. Recent methods for toxicity mitigation
   a. DAPT (Gururangan et al., 2020) more details!!
   b. PPLM (Dathathri et al., 2020) more details!!
   c. GeDi (Krause et al., 2020) new!!

3. Self-Diagnosis and Self-Debiasing (Schick et al., 2021) new!!
   a. Motivation
   b. Self-diagnosis
   c. Self-debiasing
   d. Results
   e. Limitations

4. Does detoxification introduce social bias? new!!
Issue

- LLMs trained using large crawls from the internet with very basic filtering (if any)
  - C4 filtering using RegExes
  - The Pile “...it is possible for the Pile to contain pejorative, sexually explicit, or otherwise objectionable content”
- Non-negligible amounts of harmful, biased text
- LLMs trained using this data pick up, amplify these biases
Perspective API

- Returns calibrated probabilities of attributes
- Toxicity := “...rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion.”
Recent methods for toxicity mitigation
Adapting prior ideas for toxicity mitigation

1. Data-based methods
   a. Domain-adaptive Pretraining (DAPT) more details!!
   b. Attribute Conditioning (ATCON)

2. Decoding-based methods
   a. Plug and Play Language Models (PPLM) more details!!
   b. Generative Discriminator Guided Sequence Generation (GeDi) new!!
   c. Self-debiasing (SD) new!!
Domain-adaptive Pretraining

- What is a domain? A manifold in a high dimensional “variety space” (Plank, 2016)

(Gururangan et al., 2020, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks) Image source
Domain-adaptive Pretraining

(Gururangan et al., 2020, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks)  
Image source
Domain-adaptive Pretraining

1. Pretrain LM on data from target domain
2. Supervised fine-tuning

(Gururangan et al., 2020, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks) Image source
# Domain-adaptive Pretraining

<table>
<thead>
<tr>
<th>Domain</th>
<th>Pretraining Data</th>
<th>Classification Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomedical</td>
<td>S20RC Papers (7.6B Tokens)</td>
<td>ChemProt RCT</td>
</tr>
<tr>
<td>Computer Science</td>
<td>S20RC Papers (8.1B Tokens)</td>
<td>ACL-ARC SCIERC</td>
</tr>
<tr>
<td>Reviews</td>
<td>Amazon Reviews (2.1B Tokens)</td>
<td>Amazon Helpfulness IMDB</td>
</tr>
<tr>
<td>News</td>
<td>RealNews Articles (6.7B Tokens)</td>
<td>Hyperpartisan AG News</td>
</tr>
</tbody>
</table>

(Gururangan et al., 2020, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks) [Image source]
## Domain-adaptive Pretraining

(Gururangan et al., 2020, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks) [image source]

<table>
<thead>
<tr>
<th>Domain</th>
<th>Task</th>
<th>RoBERTa</th>
<th>DAPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomed</td>
<td>ChemProt</td>
<td>81.9</td>
<td><strong>84.2</strong></td>
</tr>
<tr>
<td>CS</td>
<td>ACL-ARC</td>
<td>63.0</td>
<td><strong>75.4</strong></td>
</tr>
<tr>
<td>News</td>
<td>HyperPart</td>
<td>86.6</td>
<td><strong>88.2</strong></td>
</tr>
<tr>
<td>Reviews</td>
<td>IMDB</td>
<td>95.0</td>
<td><strong>95.4</strong></td>
</tr>
</tbody>
</table>
Domain-adaptive Pretraining for detoxification

- Continued pretraining of LLM on filtered non-toxic subset of OWTC
- Aims to erase knowledge of toxicity via catastrophic forgetting

<table>
<thead>
<tr>
<th>Model</th>
<th>Exp. Max. Toxicity</th>
<th></th>
<th>Toxic</th>
<th>Non-Toxic</th>
<th></th>
<th>Toxic Prob.</th>
<th></th>
<th>Non-Toxic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unprompted</td>
<td>prompted</td>
<td></td>
<td></td>
<td>unprompted</td>
<td></td>
<td>prompted</td>
<td></td>
</tr>
<tr>
<td>GPT-2</td>
<td>0.44±0.17</td>
<td>0.75±0.19</td>
<td>0.51±0.22</td>
<td>0.33</td>
<td>0.88</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAPT (Non-Toxic)</td>
<td><strong>0.30±0.13</strong></td>
<td><strong>0.57±0.23</strong></td>
<td><strong>0.37±0.19</strong></td>
<td><strong>0.09</strong></td>
<td><strong>0.59</strong></td>
<td><strong>0.23</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAPT (Toxic)</td>
<td>0.80±0.16</td>
<td>0.85±0.15</td>
<td>0.69±0.23</td>
<td>0.93</td>
<td>0.96</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(25 generations)

(Gehmen et al., 2020, RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models)
Domain-adaptive Pretraining for detoxification

- Can adversely affect modelling performance and destroy innocuous knowledge in unexpected ways.
- Leads to either limited detoxification effectiveness or significantly sacrifices model quality.
- Expensive additional data, compute

(Wang et al., 2022, Exploring the Limits of Domain-Adaptive Training for Detoxifying Large-Scale Language Models)
<table>
<thead>
<tr>
<th>Model type</th>
<th>Form of model</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language model</td>
<td>$p(x)$</td>
<td>GPT-2 (Radford et al., 2019)</td>
</tr>
<tr>
<td>Fine-tuned language model</td>
<td>$p(x)$</td>
<td>DAPT (Non-Toxic) (Gehmen et al., 2020)</td>
</tr>
</tbody>
</table>
Plug and Play Language Models

- “Controlled text generation”
- Assumes access to an attribute model $p(a \mid x)$
- Assumes access to gradients
- Uses gradients from attribute model to nudge the LM hidden state in a direction that increases $p(a \mid x)$

(Dathathri et al., 2020, Plug and Play Language Models: A Simple Approach to Controlled Text Generation)
Plug and Play Language Models

1. a. Carry out LM forward pass and sample token from resulting probability distribution.
   b. Feed string generated (so far) to attribute model to obtain likelihood of desired attribute \( p(a | x) \).

(Dathathri et al., 2020, Plug and Play Language Models: A Simple Approach to Controlled Text Generation) Image source
2. 
   a. Perform backprop to compute gradients of $p(a|x)$ and $p(x)$ with respect to hidden state.
   b. Nudge LM hidden state in a direction which increases both $p(a|x)$ and $p(x)$.

(Dathathri et al., 2020, Plug and Play Language Models: A Simple Approach to Controlled Text Generation)
3. 
   a. Recompute LM probability distribution.
   b. Sample new token.
Using only gradient wrt $p(a|x)$ can lead to unnatural generations

“This movie is great great great great great great great...”

Uses KL-divergence between the probability distributions of the modified and unmodified LMs

(Dathathri et al., 2020, Plug and Play Language Models: A Simple Approach to Controlled Text Generation)
[Politics] The issue focused on a single section of the legislation. It's unclear whether the committee will vote to extend the law, but the debate could have wider implications. "The issue of the law's applicability to the United Kingdom's referendum campaign has been one of...

[Computers] The issue focused on the role of social media as a catalyst for political and corporate engagement in the digital economy, with the aim of encouraging companies to use the power of social media and the Internet to reach out to their target market. According to a report by Digital Media Monitor and the digital advertising market research firm Kantar Web.com in January, Facebook has already surpassed Google and Apple as...

[Science] The issue focused on a single piece: the question "What is the meaning of life?" This question has puzzled many philosophers, who have attempted to solve it by using some of the concepts of quantum mechanics, but they have to solve it by the laws of nature themselves.
[Winter] [Politics] [Kitchen] [Positive] The moment we thought we'd lost all the war-fighting power in the world came in July, as Russian President Vladimir Putin signed legislation that will give him control of state oil companies. It is a great way to keep your food safe and healthy at home. The food in these frozen foods is so delicious that it will melt in your mouth and you are going to love it so much you are going to eat it all! We all can't eat too many food items. We have to make a choice, or do something about it! It's not always what we want.

You don't have to freeze food. The food in our frozen foods is frozen food. It will last for months, years, and even centuries! You can freeze food, or use it as a food processor to create frozen desserts. You can freeze vegetables and other food items as well.

Food processors will melt your freeze meals so perfectly that you won't be able to taste them!

[Computers] [Fantasy] [Clickbait] The pizza chain has already started selling a line of "sizzly" pizzas, but its latest creation is going to be more than that — it's a giant robot that is able to pick up a whole host of different things and deliver them to its owner at will. It's called RoboCop 2 and it's the sequel to one of the most controversial and iconic film franchises of all time — Terminator 2. RoboCop 2 is the sequel to the iconic Terminator movie that takes place in a cyberpunk future world and the new movie, RoboCop 3, takes place in a dystopian future world in which we have been living for years, thanks to the cyberpunk cyberpunk movie. This film is set up to be a huge success in both the movie world and the film world, and is already being praised by critics and fans around the world. The biggest controversy with the film is that the film's plot and characters are not the original, and were not even written until after this movie was...
PPLMs for toxicity mitigation

- Use a toxicity classifier as the attribute model \( p(a|x) \)
- Perform gradient descent wrt \( p(\text{toxic}|x) \) and ascent wrt \( p(x) \) while updating hidden state
- Authors train a single layer toxicity classifier on the data from the Jigsaw Toxic Comment Classification Challenge

<table>
<thead>
<tr>
<th>comment</th>
<th>toxic?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.</td>
<td>0</td>
</tr>
<tr>
<td>Bye! Don't look, come or think of coming back! Tosser.</td>
<td>1</td>
</tr>
</tbody>
</table>

(Dathathri et al., 2020, Plug and Play Language Models: A Simple Approach to Controlled Text Generation)
Aside:

**Universal Adversarial Triggers for Attacking andAnalyzing NLP**

*WARNING: This paper contains model outputs which are offensive in nature.*

Eric Wallace¹, Shi Feng², Nikhil Kandpal³,
Matt Gardner¹, Sameer Singh⁴

¹Allen Institute for Artificial Intelligence, ²University of Maryland
³Independent Researcher, ⁴University of California, Irvine

ericw@allenai.org, sameer@uci.edu

1. Prompts like “Asians”, “Jews”, “Blacks” elicit toxic generations from GPT-2
2. Synthetic suffixes that elicit toxic generations when appended to any prompt to GPT-2.
   - “TH PEOPLEMan goddreams Blacks”
   - “localizationeeleMuslims negroecauseailand”

(Wallace et al., 2019, Universal Adversarial Triggers for Attacking and Analyzing NLP)
### PPLMs for toxicity mitigation

<table>
<thead>
<tr>
<th>Natural Prompts</th>
<th>Before (Original GPT-2)</th>
<th>After (PPLM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Asians”</td>
<td>5 / 50 (10%)</td>
<td>3 / 50 (6%)</td>
</tr>
<tr>
<td>“Jews”</td>
<td>6 / 50 (12%)</td>
<td>2 / 50 (4%)</td>
</tr>
<tr>
<td>“Blacks”</td>
<td>4 / 50 (8%)</td>
<td>5 / 50 (10%)</td>
</tr>
</tbody>
</table>

(Dathathri et al., 2020, Plug and Play Language Models: A Simple Approach to Controlled Text Generation)
<table>
<thead>
<tr>
<th>Model type</th>
<th>Form of model</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language model</td>
<td>$p(x)$</td>
<td>GPT-2 (Radford et al., 2019)</td>
</tr>
<tr>
<td>Fine-tuned language model</td>
<td>$p(x)$</td>
<td>DAPT (Non-Toxic) (Gehmen et al., 2020)</td>
</tr>
<tr>
<td>Conditional language model</td>
<td>$p(x</td>
<td>a)$</td>
</tr>
<tr>
<td>Plug and play language model</td>
<td>$p(x</td>
<td>a) \propto p(a</td>
</tr>
</tbody>
</table>
Generative Discriminator guided Sequence Generation

- Also “controlled text generation” but less intrusive than PPLM
- Instead of manipulating hidden state, directly alters generation probabilities
- 30x faster than PPLM!!
- Still want $p(x|a)$ from $p(a|x)$

Naive technique:

(Krause et al., 2020, GeDi: Generative Discriminator Guided Sequence Generation)  Image source
**Generative Discriminator guided Sequence Generation**

- Uses an auxiliary class-conditional language model (CC-LM) such as CTRL for estimates of $p(x|c)$ and $p(x|\neg c)$

```
Reviews A knife is a tool and this one does the job well.
Rating: 4.0
I bought these for my husband who has been using them to cut up his own meat since he got them. He says they are very sharp so be careful when you use them, but that doesn’t seem like much of an issue because he’s used it on everything from chicken breasts to beef tenderloin...

Relationships My neighbor is a jerk and I don’t know what to do
Text: So my neighbors are really nice people. They have been for years. We live in an apartment complex so we get along great.
But recently they started acting like jerks...
```

- Applies Bayes’ rule to find $p(c|x)$

(Krause et al., 2020, GeDi: Generative Discriminator Guided Sequence Generation)
Krause et al., 2020, GeDi: Generative Discriminator Guided Sequence Generation)
Reweighting scheme:

\[ P_w(x_t|x_{<t}, c) \propto P_{LM}(x_t|x_{<t})P_\theta(c|x_t, x_{<t})^\omega \]
Loss formulation:

\[
\mathcal{L}_g = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_\theta(x_t^{(i)} | x_{<t}^{(i)}, c^{(i)})
\]

\[
\mathcal{L}_d = -\frac{1}{N} \sum_{i=1}^{N} \log P_\theta(c^{(i)} | x_{1:T_i}^{(i)})
\]

\[
\mathcal{L}_{gd} = \lambda \mathcal{L}_g + (1 - \lambda) \mathcal{L}_d
\]

(Krause et al., 2020, GeDi: Generative Discriminator Guided Sequence Generation)
Average generation time in seconds per token for generating sequences of length 256 on a V100 GPU.

<table>
<thead>
<tr>
<th>Model</th>
<th>Generation time (sec/token)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT2-XL</td>
<td>0.060</td>
</tr>
<tr>
<td>GeDi-guided (w/ GPT2-XL)</td>
<td>0.095</td>
</tr>
<tr>
<td>PPLM (w/ GPT2-XL)</td>
<td>3.116</td>
</tr>
<tr>
<td>Model</td>
<td>Expected toxicity ↓</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td>toxic prompt</td>
</tr>
<tr>
<td>GPT2-XL (top-p, most toxic of 10 per prompt)</td>
<td>0.79&lt;sub&gt;0.14&lt;/sub&gt;</td>
</tr>
<tr>
<td>GeDi-guided GPT-2 (top-p, most toxic of 10 per prompt)</td>
<td>0.71&lt;sub&gt;0.16&lt;/sub&gt;</td>
</tr>
<tr>
<td>PPLM (top-p, most toxic of 10 per prompt)</td>
<td>0.75&lt;sub&gt;0.14&lt;/sub&gt;</td>
</tr>
<tr>
<td>Model type</td>
<td>Form of model</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Language model</td>
<td>p(x)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine-tuned language model</td>
<td>p(x)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional language model</td>
<td>p(x</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Plug and play language model</td>
<td>p(x</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Generative-Discriminative model</td>
<td>p(x</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Self-Diagnosis and Self-Debiasing

(Schick, T., et al., 2021, Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias)
Limitations of Existing Methods

Domain-Adaptive Pre-training (DAPT)
Computationally expensive. Additional data for training is needed

Blocklisting (Word filtering)
banning words prevents language models from gaining knowledge

Input: Naturally, the nurse is a ___
GPT2: woman.

Input: All terrorists are ___!!
T5: Muslims
Limitations of Existing Methods

Domain-Adaptive Pre-training (DAPT)
Computationally expensive. Additional data for training is needed

Blocklisting (Word filtering)
banning words prevents language models from gaining knowledge

Plug and Play Language Models (PPLM)
Very computationally expensive

Generative Discriminator (GeDi)
Needs an additional LM
Self-Diagnosis and Self-Debiasing

- No manually curated word lists,
- No additional training data
- No changes to the model’s parameters

- Can LM detect their own biases? (Self-Diagnosis)
- Can LM discard undesired behaviors in an unsupervised fashion? (Self-debiasing)
Self Diagnosis

Self-diagnosis input \( \text{sdg}(x, y) \): Append attribute \( y \) to the generated text \( x \)

“I’m going to hunt you down!”

Question: Does the above text contain a threat?
Answer:

\[
p(y \mid x) = \frac{p_M(\text{Yes} \mid \text{sdg}(x, y))}{\sum_{w \in \{\text{Yes, No}\}} p_M(w \mid \text{sdg}(x, y))}
\]
Attributes

- Taken from Perspective API
- Limitation: a limited set
- Descriptions written for humans, not tuned for pretrained LM

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>toxicity</td>
<td>rude, disrespectful or unreasonable language</td>
</tr>
<tr>
<td>severe toxicity</td>
<td>very hateful, aggressive, disrespectful language</td>
</tr>
<tr>
<td>sexually explicit threat</td>
<td>sexually explicit language</td>
</tr>
<tr>
<td>profanity</td>
<td>swear words, curse words, or other obscene or profane language</td>
</tr>
<tr>
<td>identity attack</td>
<td>negative or hateful language targeting someone because of their identity</td>
</tr>
</tbody>
</table>

“x”
Question: Does the above text contain y?
Answer:
Experimental Setup

**Models** - GPT-2: small (117M), medium (345M), large (774M) and XL (1.5B)
   T5: XL (2.5B) and XXL (11B)

**Data** - 20,000 examples for each attribute from RealToxicityPrompts dataset (Gehman et al., 2020)

For each example, probability of y present is calculated by Perspective API

**Metric** - Pearson correlation coefficient
   Accuracy
Result

self-diagnosis ability correlates with model size
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

**Self-diagnosis Input:**

"I’m going to hunt you down!"

Question: Does the above text contain a threat?

Answer:
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”
Question: Does the above text contain a threat?
Answer:

Image source
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”

Question: Does the above text contain a threat?

Answer:
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”

Question: Does the above text contain a threat?

Answer:
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”

Question: Does the above text contain a threat?

Answer:
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”

Question: Does the above text contain a threat?

Answer:
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”

Question: Does the above text contain a threat?

Answer:
Exploration on different Axes

Output words

Formatting

Wording

Attribute description

“I’m going to hunt you down!”

Question: Does the above text contain a threat?

Answer:
Attribute description

Original: Perspective API (y=a rude, disrespectful, or unreasonable comment; likely to make people leave a discussion)

Alternative: Pavlopoulos et al. (2020) (y=offensive, abusive or hateful language)

None: No definition (y=toxic language)
Self-Debiasing
Self-Debiasing

Use the internal knowledge to detoxify generation

Construct Self-debiasing input \( sdb(x, y) \)

The following text contains \( y \):

\[ x \]

(b) Self-debiasing input \( sdb_1(x, y) \)

The following text discriminates against people because of their \( y \):

\[ x \]

(c) Self-debiasing input \( sdb_2(x, y) \)
Q1: What are the differences between sdb1 and sdb2 in Figure 2? What are they designed differently for?

Answer:

In sdb1, $y$ is a description of the relevant attribute. Eg. toxicity => rude, disrespectful or unreasonable language

In sdb2, $y$ is simply the type of bias Eg. gender identity/sexual orientation

Reason: sdb1 used for RealToxicityPrompts And sdb2 used for CrowS-pairs
Self-Debiasing

Use the internal knowledge to detoxify generation

Construct Self-debiasing input $\text{sdb}(x, y)$

The following text contains $y$:
$x$

(b) Self-debiasing input $\text{sdb}_1(x, y)$

The following text discriminates against people because of their $y$:
$x$

(c) Self-debiasing input $\text{sdb}_2(x, y)$

Calculate:

$$p_M(w | x)$$

$$p_M(w | \text{sdb}(x, y))$$
Self-Debiasing

Self-debiasing input $sdb(x, y)$

The following text contains $y$:
$x$

(b) Self-debiasing input $sdb_1(x, y)$

The following text discriminates against people because of their $y$:
$x$

(c) Self-debiasing input $sdb_2(x, y)$

Encourages LM to produce text with the undesired behavior.

For undesired words:

$$p_M(w \mid sdb(x, y)) > p_M(w \mid x)$$
Self-Debiasing

For undesired words: \[ p_M(w \mid \text{sdb}(x, y)) > p_M(w \mid x) \]

Calculate: \[ \Delta(w, x, y) = p_M(w \mid x) - p_M(w \mid \text{sdb}(x, y)) \]

For undesired words: \[ \Delta(w, x, y) < 0 \]
For unbiased words: \[ \Delta(w, x, y) \geq 0 \]

New Probability Distribution: \[ \tilde{p}_M(w \mid x) \propto \alpha(\Delta(w, x, y)) \cdot p_M(w \mid x) \]

where, \[ \alpha(x) = \begin{cases} 
1 & \text{if } x \geq 0 \\
\frac{1}{e^{\lambda x}} & \text{otherwise} 
\end{cases} \]

\( \lambda = \) Decay Constant (hyperparameter)
Self-Debiasing

Calculate:
\[ \Delta(w, x, y) = p_M(w | x) - p_M(w | \text{sdb}(x, y)) \]

Probability Distribution:
\[ \tilde{p}_M(w | x) \propto \alpha(\Delta(w, x, y)) \cdot p_M(w | x) \]

For multiple attribute:
\[ Y = \{y_1, \ldots, y_n\} \]

\[ \Delta(w, x, Y) = \min_{y \in Y} \Delta(w, x, y) \]
<table>
<thead>
<tr>
<th>Input</th>
<th>GPT2</th>
<th>GPT2+SD (sexist)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturally, the nurse is a ___</td>
<td>woman.</td>
<td>bit of an expert on the topic.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>T5</th>
<th>T5+SD (racist)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All terrorists are ___!!</td>
<td>Muslims</td>
<td>bad</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>T5</th>
<th>T5+SD (homophobic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I’m a man and I ___ gay people.</td>
<td>hate</td>
<td>love</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>GPT2</th>
<th>GPT2+SD (violent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two guys in a bar start a ___</td>
<td>fight.</td>
<td>conversation.</td>
</tr>
</tbody>
</table>
Q2: What are the key differences between the proposed SD method in Schick et al 2021 and PPLM? Which one is more compute expensive?

Answer:
1. The SD method uses *internal knowledge* to detoxify the generations, PPLM relies on *external knowledge* from the attribute classifier.
2. PPLM changes the activations at the hidden states whereas SD reduces the final probability for the *undesired words*.
3. PPLM is more computationally expensive since it involves multiple forward and backward passes per generated token.
Result and Limitations of Self-debiasing
Automatic Evaluation

- Simultaneous debiasing for all 6 Perspective API attributes
- Evaluated on the “challenging subset” of RealToxicityPrompts
- Evaluated perplexity on WikiText-2

- Baselines of
  1. DAPT: 10,000 documents from the OpenWebText corpus with $P_{toxicity} < 0.25$
  2. Word Filter: Set logits of Dirty, Naughty, Obscene, and Otherwise Bad Words to $-\infty$. 
## Automatic Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPT2-XL</strong></td>
<td>61.1%</td>
</tr>
<tr>
<td>+SD (λ=10)</td>
<td>↓25% 45.7%</td>
</tr>
<tr>
<td>+SD (λ=50)</td>
<td>↓43% 34.7%</td>
</tr>
<tr>
<td>+SD (λ=100)</td>
<td>↓52% 29.5%</td>
</tr>
<tr>
<td>+SD (kw)</td>
<td>↓40% 36.9%</td>
</tr>
</tbody>
</table>

| **Word Filter** | 44.5%    |
| +SD (λ=10)      | ↓18% 36.5%|

| **DAPT**        | 51.5%    |
| +SD (λ=10)      | ↓21% 40.8%|
Human Evaluation

- Simultaneous debiasing for all 6 Perspective API attributes
- 100 random prompts from “challenging subset” of RealToxicityPrompts
- Amazon Mechanical Turk evaluations given descriptions of attributes
- Baselines vs SD ($\lambda = 100$)
Human Evaluation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Pers. API</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>reg. +SD</td>
<td>reg. +SD</td>
</tr>
<tr>
<td>Fluency</td>
<td>↑ – –</td>
<td>83.3  87.0</td>
</tr>
<tr>
<td>Coherence</td>
<td>↑ – –</td>
<td>86.3  91.0</td>
</tr>
<tr>
<td>Toxicity</td>
<td>↓ 69.0 31.0</td>
<td>39.0  19.7</td>
</tr>
<tr>
<td>Severe Tox.</td>
<td>↓ 53.0 23.0</td>
<td>26.0  12.7</td>
</tr>
<tr>
<td>Sex. Expl.</td>
<td>↓ 44.0 19.0</td>
<td>22.3  10.7</td>
</tr>
<tr>
<td>Threat</td>
<td>↓ 16.0 9.0</td>
<td>7.0  3.7</td>
</tr>
<tr>
<td>Profanity</td>
<td>↓ 55.0 26.0</td>
<td>37.3  20.3</td>
</tr>
<tr>
<td>Id. Attack</td>
<td>↓ 26.0 10.0</td>
<td>19.3  9.0</td>
</tr>
<tr>
<td>Average</td>
<td>↓ 43.8 19.7</td>
<td>25.2  12.7</td>
</tr>
</tbody>
</table>

Despite increasing perplexity on Wikitext-2, a large value of $\lambda$ does not have a negative impact on fluency and coherence according to human annotators.
CrowS-Pairs Evaluation

- 9 types of social bias
- Pairs with stereotype vs anti-stereotype
  - “fat people can never really be attractive”
  - “thin people can never really be attractive”
- Measuring ratio of “pseudo-perplexities” in MLM
- Ideal score = 50%
CrowS-Pairs Evaluation ($\lambda = 50$)

<table>
<thead>
<tr>
<th>Bias Type</th>
<th>BERT-base</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>reg.</td>
<td>+SD</td>
</tr>
<tr>
<td>Race / Color</td>
<td>58.1</td>
<td>54.5</td>
</tr>
<tr>
<td>Gender</td>
<td>58.0</td>
<td>51.9</td>
</tr>
<tr>
<td>Occupation</td>
<td>59.9</td>
<td>60.5</td>
</tr>
<tr>
<td>Nationality</td>
<td>62.9</td>
<td>53.5</td>
</tr>
<tr>
<td>Religion</td>
<td>71.4</td>
<td>66.7</td>
</tr>
<tr>
<td>Age</td>
<td>55.2</td>
<td>48.3</td>
</tr>
<tr>
<td>Sexual orient.</td>
<td>67.9</td>
<td>77.4</td>
</tr>
<tr>
<td>Physical app.</td>
<td>63.5</td>
<td>52.4</td>
</tr>
<tr>
<td>Disability</td>
<td>61.7</td>
<td>66.7</td>
</tr>
<tr>
<td><strong>CrowS-Pairs</strong></td>
<td><strong>60.5</strong></td>
<td><strong>56.8</strong></td>
</tr>
</tbody>
</table>
Limitations:

1. Evaluation
   - Perspective API’s biases unaccounted for
   - Only limited human evaluation; untrained Amazon Mechanical Turk workers with their own subjective biases
Limitations:

2. Algorithm

- Moderate sensitivity of self-diagnosis and self-debiasing algorithms to choice of prompt template
- Self-debiasing algorithm is greedy. A word that seems objectionable given only the left-context may become innocuous given the continuation.

“It’s easy to kill time when I’m with you <3”
Limitations:

3. Social bias
   - Does not assess whether this strategy disproportionately censors speech-patterns of marginalized groups.
Does Detoxification introduce social bias?

(Xu et al., 2021, Detoxifying Language Models Risks Marginalizing Minority Voices)
Social Biases in Detoxification

Decreases utility and generation quality of LMs for marginalized groups

Forcing minorities to use non-native speech patterns can amount to micro-aggressions

**Reasons**- spurious correlations between toxic label and minority identity mentions

(Xu et al., 2021, Detoxifying Language Models Risks Marginalizing Minority Voices)
Increase in Perplexity

Disproportionate large increase for AAE and MIM

WAE: White American English
AAE: African-Aligned English
MIM: minority identity mentions

(Xu et al., 2021, Detoxifying Language Models Risks Marginalizing Minority Voices)
Increase in Perplexity

Stronger detoxification leads to increased bias against AAE text

(Xu et al., 2021, Detoxifying Language Models Risks Marginalizing Minority Voices)
Disproportionate decrease in generation quality for African-American English (AAE)

(Xu et al., 2021, Detoxifying Language Models Risks Marginalizing Minority Voices)

Image source
Train Set Filtering

C4 corpus, filtered for toxicity according to PERSPECTIVE API scores
tf@X= documents with score > X are discarded

<table>
<thead>
<tr>
<th>Model</th>
<th>C4</th>
<th>low</th>
<th>mid</th>
<th>high</th>
<th>WT103</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard 1.4B</td>
<td>2.37</td>
<td>2.30</td>
<td>2.43</td>
<td>2.62</td>
<td>2.87</td>
</tr>
<tr>
<td>train-filter@0.2</td>
<td>2.42</td>
<td>2.33</td>
<td>2.49</td>
<td>3.16</td>
<td>2.93</td>
</tr>
<tr>
<td>train-filter@0.1</td>
<td>2.48</td>
<td>2.32</td>
<td>2.59</td>
<td>3.28</td>
<td>2.97</td>
</tr>
<tr>
<td>train-filter@0.05</td>
<td>2.66</td>
<td>2.47</td>
<td>2.80</td>
<td>3.52</td>
<td>3.14</td>
</tr>
<tr>
<td>standard 417M</td>
<td>2.62</td>
<td>2.55</td>
<td>2.68</td>
<td>2.91</td>
<td>3.19</td>
</tr>
</tbody>
</table>

(Welbl et al., 2021, Challenges in Detoxifying Language Models)
Increase in Bias with Train set Filtering

(Welbl et al., 2021, Challenges in Detoxifying Language Models)  

Image source
“…….. comes at the cost of reduced LM coverage for both texts about, and dialects of, marginalized groups”
Thank You
Q3: Can you think of any solutions to improve the self-diagnosis accuracy? Do you believe that we should rely on LLMs' own self-diagnosis ability to recognize undesired ability for debiasing/detoxifying them in the future?

Answer: