Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness

Fatemehsadat Mireshghallah, Taylor Berg-Kirkpatrick
fatemeh@ucsd.edu, tberg@ucsd.edu

EMNLP 2021
Text Style Can Bias Our Assumptions about the Author

it was soooo fricken hilarious.

Text style can lead to assumptions on the author’s:

• Age
• Gender
• Identity

Such assumptions can affect future decisions based on the text.
Text Style Can Bias Our Assumptions about the Author

These biases can have significant impact on objectivity in high stakes situations.
Landscape of Fairness in NLP

• De-biasing word embeddings: Ravfogel et al. (2020), Kaneko and Bollegala (2019), Shin et al. (2020)

• De-biasing model representations and encodings: Elazar and Goldberg (2018); Barrett et al. (2019); Wang et al. (2021)

• ML methods of modifying learning algorithms/inference procedures to get more fair outcomes: Agarwal et al., 2018; Madras et al., 2018; Zafar et al., 2017…
Prior work: A4NT - Hide attributes by imitation

Hide sensitive attributes by translating text from one attribute domain to another.
Prior work: A4NT - Hide attributes by imitation

Limitations of A4NT:
• Imitating attributes does not eliminate the bias, it just shifts it!
• Supporting sensitive attributes with more than two is non-trivial

it was soooo fricken hilarious.

it was undisputedly hilarious.
Style-pooling: Central Notion of Style

- Let’s assume we have a Dataset $D$, of blogs, which has annotations for attribute $A$, age group of each blog’s author (Teen, Young Adult and Adult).
- Based on $A$, we can partition the data intro three domains, $D_1, D_2, D_3$, corresponding to teens, young adults and adults.
- **Goal:** Transfer observed text utterance to a semantically equivalent text utterance with obfuscated style.

\[
\begin{align*}
   &x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
\end{align*}
\]

- $D_1$: Teen
- $D_2$: Young Adult
- $D_3$: Adult

Ex:  

- Input: $x_1$: it was so hilarious.  
- Output: $y_1$: Now I was intrigued.

Ex:

- Input: $x_2$: it was soooo fricken hilarious.  
- Output: $y_6$: Now I was intrigued.
Style-pooling: Central Notion of Style

- Treat the obfuscated text as latent variables in a generative model.
- Treat the training data as observed variables.
- For each observed variable \( x \) there will be a corresponding latent variable \( y \).

\[ x_1 \rightarrow y_1 \]
\[ x_2 \rightarrow y_2 \]
\[ x_3 \rightarrow y_3 \]
\[ x_4 \rightarrow y_4 \]
\[ x_5 \rightarrow y_5 \]
\[ x_6 \rightarrow y_6 \]

\( D_1: \text{Teen} \)
\( D_2: \text{Young Adult} \)
\( D_3: \text{Adult} \)

Observed training data from each domain

Unseen obfuscated data
Style-pooling: Central Notion of Style

- We assume each observed sentence $x$ is generated as follows:
  1. $y \sim p_{prior}$

$$p_{Prior}(y) = \frac{1}{3} p_1 + \frac{1}{3} p_2 + \frac{1}{3} p_3$$
Style-pooling: Central Notion of Style

- We assume each observed sentence $x$ is generated as follows:
  1. $y \sim p_{\text{prior}}$
  2. $x \sim p(x|y; \theta^d_{y \rightarrow x})$, where $\theta^d_{y \rightarrow x}$ are the parameters for the seq2seq transduction model, for decoding to $x$’s domain, $d$.

\[ p_{\text{Prior}}(y) = \frac{1}{3} p_1 + \frac{1}{3} p_2 + \frac{1}{3} p_3 \]
Learning

• To learn the parameters $\theta_{y\rightarrow x}^d$, we need to maximize the likelihood of the observed data, which can be parameterized as:

$$
\log p(X^1, X^2, X^3; \theta_{y\rightarrow x}^1, \theta_{y\rightarrow x}^2, \theta_{y\rightarrow x}^3) = \log \sum_Y \prod_{i=1}^N p(x_i | y_i; \theta_{y\rightarrow x}^{d(i)}) p_{\text{prior}}(y_i)
$$

• The summation over the latent representation $Y$ is intractable.

$$
P_{\text{Prior}}(y) = \frac{1}{3} p_1 + \frac{1}{3} p_2 + \frac{1}{3} p_3
$$
Learning

• We use a seq2seq inference network*, parameterized by $\phi_{x\rightarrow y}$ to learn $q(y|x)$, which approximates the posterior $p(y|x)$.

• We use an amortized inference scheme, the same approach in VAEs to solve this problem.

* He et al., A probabilistic formulation of unsupervised text style transfer, ICLR 2020
Learning

We want to maximize log likelihood of the data

\[
\log p(X^1, X^2, \ldots, X^M; \theta^{1}_{y \rightarrow x}, \theta^{2}_{y \rightarrow x}, \ldots, \theta^{M}_{y \rightarrow x})
\]

Intractable

\[
\geq \mathcal{L}_{ELBO}(X^1, X^2, \ldots, X^M; \theta^{1}_{y \rightarrow x}, \theta^{2}_{y \rightarrow x}, \ldots, \theta^{M}_{y \rightarrow x}, \phi_{x \rightarrow y})
\]

We use a tractable lower bound on the data likelihood (evidence lower bound, ELBO)

\[
= \sum_{i}^{N} \left[ \mathbb{E}_{q(y_i|x_i; \phi_{x \rightarrow y})} \left[ \log p(x_i|y_i; \theta^{d(i)}_{y \rightarrow x}) \right] - D_{KL}(q(y_i|x_i; \phi_{x \rightarrow y}) || p_{prior}(y_i)) \right]
\]

Reconstruction Likelihood:

Latent text \( y \) is back-translated to \( x \) correctly

KL Regularizer:

Distribution of latent text \( y \) is close to the language model prior

* He et al., A probabilistic formulation of unsupervised text style transfer, ICLR 2020
Inference

\[ D_0 \text{ Obfuscated} \]

\[ y \]

\[ \text{it was so hilarious.} \]

\[ q(y|x; \phi_{x\rightarrow y}) \quad \text{Enc} \quad \text{Dec} \quad p(x|y; \theta_{y\rightarrow x}) \]

\[ x \]

\[ \text{it was sooo fricken hilarious.} \]

\[ D_1: \text{Teen} \]
Style De-boosting

- To further remove (de-boost) each domain’s style (on a word level), we propose style de-boosting.
- Style de-boosting de-incentivizes the use of words whose presence might hint at a particular sensitive attribute.

\[ s_w = \frac{\max(f_{w1}^D, f_{w2}^D, ..., f_{wM}^D) - \min(f_{w1}^D, f_{w2}^D, ..., f_{wM}^D)}{\max(f_{w1}^D, f_{w2}^D, ..., f_{wM}^D)} \]

\[ p(y_{i,t} | y_{i,<t}, x_i) \propto \text{softmax}(L_{i,t} - \gamma * S) \]
### Experimental Setup: Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
</table>
| Synthetic Yelp | 400k reviews
3 Domains based on **synthetic misspellings**                         |
| Blogs        | 3.3 million sentences of microblogs
2 and 3 Domains based on author’s **age**.                              |
| DIAL         | Twitter dataset with 59 Million Tweets
2 Domains based on binary **dialect** annotations                      |
| Twitter      | PAN16 Twitter dataset with 436 users
2 Domains based on **age**.                                              |
Experimental Setup: Metrics

Attribute Classification
- Classifier Accuracy (50% ideal)
- Classifier Entropy (1 is ideal)

Text Quality
- Back-Translation (BT) accuracy
- GPT-2 PPL
- BLEU Score
- GLEU Score
- Lexical Diversity

Fairness
- True Positive Rate (TPR)-GAP of a downstream classifier:

$$TPR_{a,y} = P(\hat{Y} = y | A = a, Y = y)$$
$$GAP_{TPR} = TPR_{a,y} - TPR_{a',y}$$
List of experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>In the Presentation</th>
<th>In the Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogs (2 domain)</td>
<td>A4NT, One-LM</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Blogs (3 domain)</td>
<td>One-LM</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>DIAL</td>
<td>PATR</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Synthetic Yelp</td>
<td>One-LM</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Twitter</td>
<td>Elazar et al.</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Blogs-Human Evaluation</td>
<td>A4NT</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Experimental Results: Comparison with A4NT on Blogs (2 domains)

- Style pooling with high de-boosting (DB) can better confuse the classifier, compared to A4NT, while preserving the fluency of the text.
Experimental Results: Comparison with A4NT on Blogs - Fairness

- For the task of debiasing job classification based on blogs, as we increase the de-boosting, the TPR-gap (fairness) metric decreases, which means improved fairness.
- For this task our method is significantly better than A4NT, especially in terms of preserving utility.
Another Prior: Union

• One limitation of our average pooling prior is that it acts as majority voting between styles, which could remove stylistic features belonging to minority groups.
• To mitigate this, we study the possibility of a union prior, which selects the minimum score of all language models.

\[ P_{\text{Union}}(y) \propto \prod_t \min(p^{D_1}(y_t|y_{<t}), \ldots, p^{D_M}(y_t|y_{<t})) \]
Union – Synthetic Yelp Data

Domains based on **synthetic misspellings:**
1. **00great00, 00this00, 00it00, 00to00, 00food00**
2. **11of11, 11place11, 11for11, 11good11, 11service11**
3. **22they22, 22are22, 22in22, 22very22, 22my22**
• The One-LM baseline is ground truth in this setup, since it was trained on correctly spelled data.
• Avg-pooling is very similar to the One-LM baseline, which demonstrates it’s effectiveness in majority voting.
• Union is successfully spreading misspellings across domains, which is the behavior expected.
Conclusions and Future Directions

• To mitigate the problem of biases created through style, we propose style-pooling, which attempts at creating a **neutral** style.

• The proposed method is capable of pooling multiple styles, unlike prior work that can only transfer between two styles.

• We put forward two possible definitions of neutral, **average** and **union** of styles.

• What could be other definitions of neutral?
Thank you!

fatemeh@ucsd.edu

Code: https://github.com/miresghallah/style-pooling/