1. Introduction

**Background:** Written text often provides clues to identify the author, their gender, age, and other important attributes. As a result, the authorship of training and evaluation corpora can have unforeseen consequences, including differing model performance for different user groups, as well as privacy implications. 

**Aim:** to learn un-biased representations which protect author’s attributes.

**Our contribution:** propose an approach to obscure important author characteristics at training time, such that representations learned are invariant to these attributes.

2. A Trustpilot Attacker Example

![Trustpilot Data Example]

- **Trustpilot: Trustpilot English POS tagged dataset (Hovy and Sogaard, 2015)**
- **BASELINE:** Bi-LSTM trained on Web English Treebank (Bies et al., 2012)
- Two evaluations: in-domain and out-of-domain.

### 1. TrustPilot English POS tagged dataset (Hovy and Sogaard, 2015)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-M</td>
<td>81.4</td>
</tr>
<tr>
<td>O45</td>
<td>81.2</td>
</tr>
<tr>
<td>U35</td>
<td></td>
</tr>
</tbody>
</table>

### 2. African-American Vernacular English (Jørgensen et al., 2016)

- **Three heterogeneous domains:** LYRICS, SUBTITLES and TWEETS

### 3. Model Architecture

![Model Architecture Diagram]

- **Model:** $\theta$, $\phi$, $\phi'$
- **Discriminator:** $D_i(\phi')$
- **Cross-Entropy Loss:** $\mathcal{L}(\hat{y}, y) = -\sum \hat{y}_j \log(y_j)$
- **Adversarial Loss:** $\mathcal{L}_{adv}(\theta) = -\mathbb{E}_{x \sim p(x)} [\log(D_i(\phi(x), b)) + \log(1 - D_j(\phi(x), b))]$
- **Total Loss:** $\mathcal{L}(\theta) = \mathcal{L}_{adv}(\theta) + \lambda \mathcal{L}(\theta)$

### 4. POS-tagging

- **BASELINE:** word-level CNN
- **Dataset:** TrustPilot dataset derived from Hovy et al. (2015)
  - Target variable: RATING
  - Three attributes: gender (SEX binary), age (AGE binary), and location (LOC (US, UK, DE, FR)).
  - Retrieve English reviews, and resample to balance LOC.
- **Evaluation:**
  - RATING: accuracy (higher is better) as main task performance,
  - Discriminator accuracy (majority is better) as attacker.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-M</td>
<td>81.4</td>
</tr>
<tr>
<td>O45</td>
<td>81.2</td>
</tr>
<tr>
<td>U35</td>
<td></td>
</tr>
</tbody>
</table>

### 5. Sentiment Analysis

- **BASELINE:** word-level CNN
- **Dataset:** TrustPilot dataset derived from Hovy et al. (2015)
  - Target variable: RATING
  - Three attributes: gender (SEX binary), age (AGE binary), and location (LOC (US, UK, DE, FR)).
  - Retrieve English reviews, and resample to balance LOC.
- **Evaluation:**
  - RATING: accuracy (higher is better) as main task performance,
  - Discriminator accuracy (majority is better) as attacker.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-M</td>
<td>81.4</td>
</tr>
<tr>
<td>O45</td>
<td>81.2</td>
</tr>
<tr>
<td>U35</td>
<td></td>
</tr>
</tbody>
</table>

- **Our method can hide much of the personal information of users, without affecting the sentiment task performance.**

https://github.com/lrank/Robust_and_Privacy_preserving_Text_Representations