Towards Robust and Privacy-preserving Text Representations

Yitong Li, Timothy Baldwin, Trevor Cohn

yitongl4@student.unimelb.edu.au, {tbaldwin,tcohn}@unimelb.edu.au

THE UNIVERSITY OF MELBOURNE

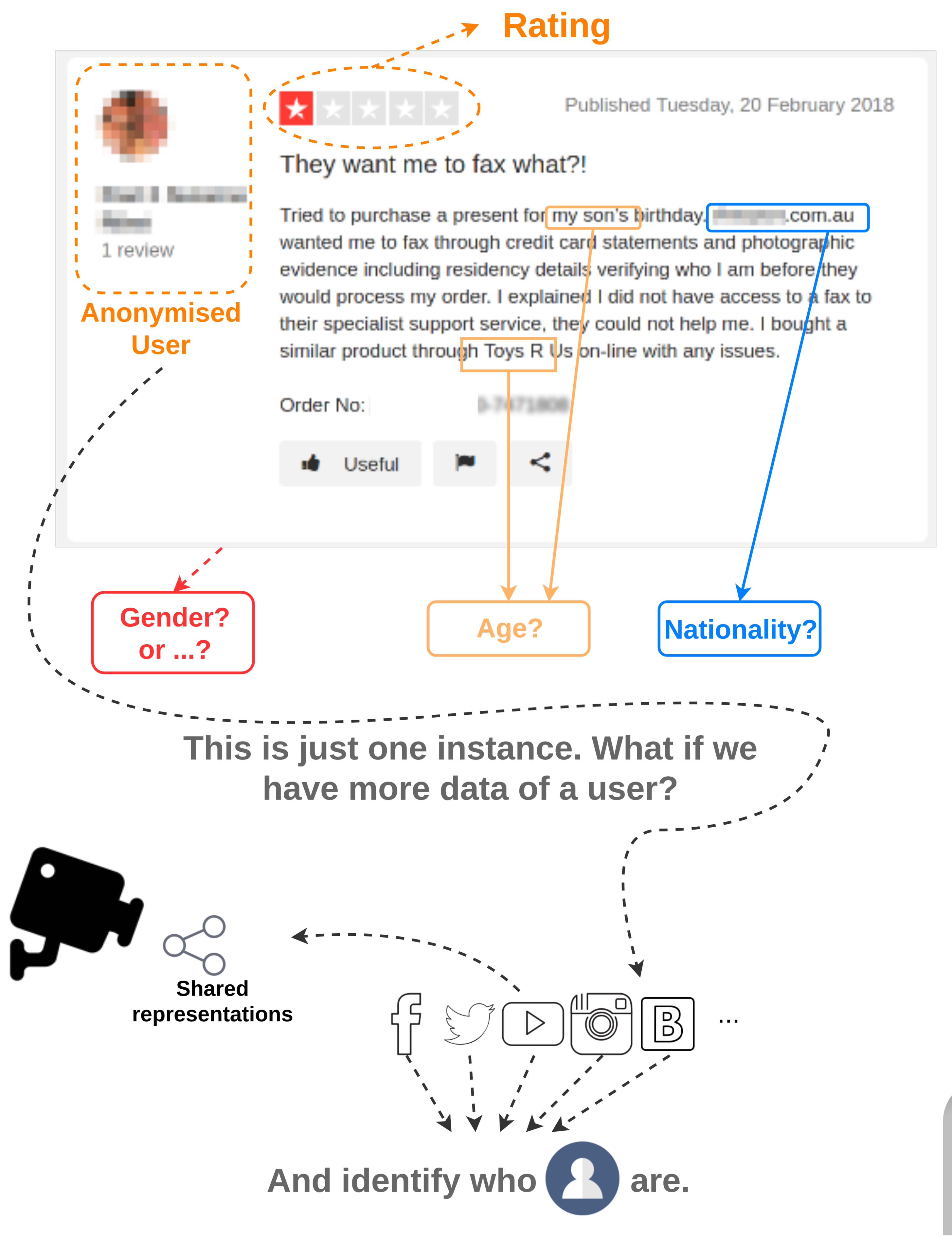
1. Introduction

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.

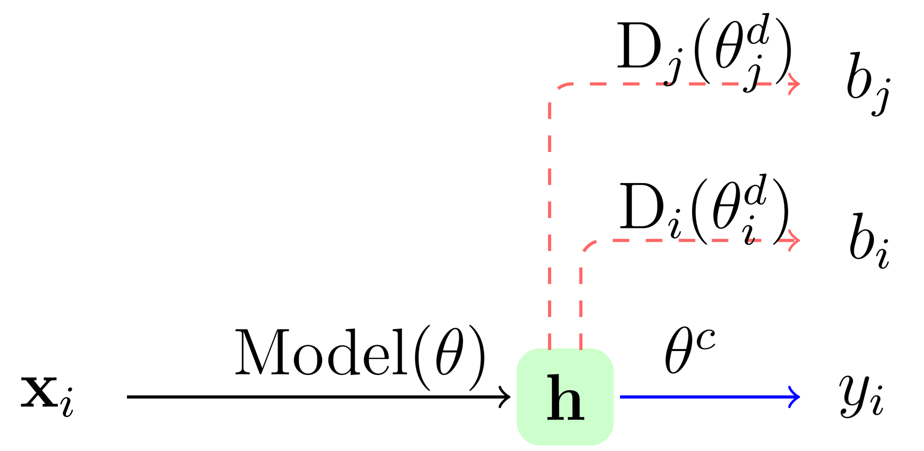
Aim: to learn un-biased representations which protect author's attributes.

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



3. Model Architecture



- $(\mathbf{x}_i, y_i)$: a training instance with two protected attributes b_i and b_j;
- $D^{\{\cdot\}}(\theta^d)$ = a discriminator, predicting the domain;
- red dashed and blue lines denote adversarial and standard loss.
- $\mathcal{X}$ = cross-entropy loss.

Formulated as:

$$\hat{\theta} = \min_{\theta_M} \max_{\{\theta_{D^i}\}_{i=1}^{N}} \mathcal{X}(\hat{\mathbf{y}}(\mathbf{x};\theta_M),\mathbf{y}) - \sum_{i=1}^{N} \lambda_i \cdot \mathcal{X}(\hat{b}(\mathbf{x};\theta_{D^i}), b_i)$$

5. Sentiment Analysis

- BASELINE: word-level CNN
- Dataset: TrustPilot dataset derived from Hovy et al. (2015)
  - Target variable: RATING $_{1-5}$
  - Three attributes: gender (SEX $_{binary}$), age (AGE $_{binary}$), and location (LOC $_{\{US,UK,GE,DE,FR\}}$).
  - 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.

	F_1		Discrim. [%]		
	dev	test	AGE	SEX	LOC
Majority class			57.8	62.3	20.0
BASELINE	41.9	40.1	65.3	66.9	53.4
ADV-AGE	**42.7**	40.1	**61.1**	65.6	41.0
ADV-SEX	42.4	39.9	61.8	62.9	42.7
ADV-LOC	42.0	**40.2**	62.2	66.8	**22.1**
ADV-all	42.0	**40.2**	61.8	**62.5**	28.1

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

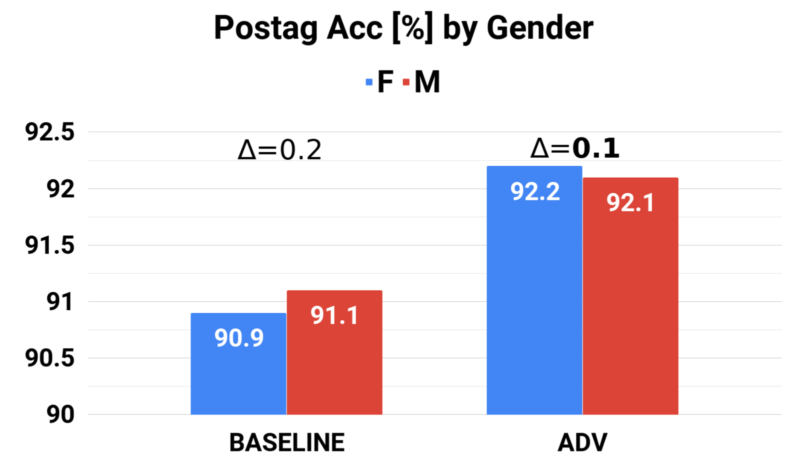
https://github.com/lrank/Robust_and_Privacy_preserving_Text_Representations

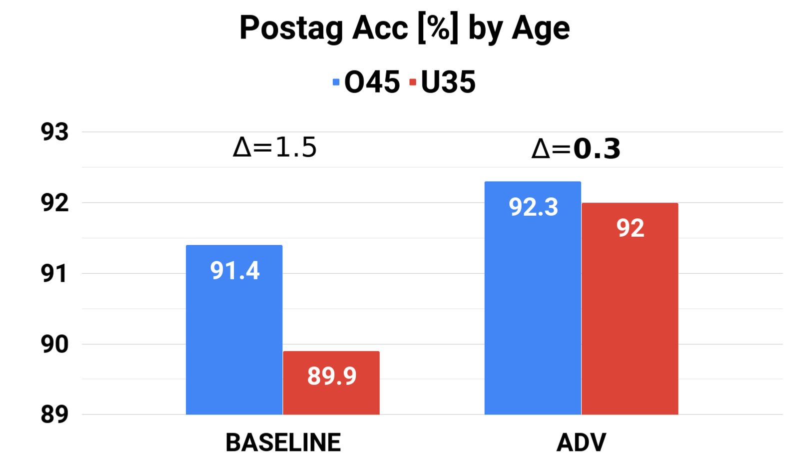
4. POS-tagging

- 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 Søgaard, 2015)

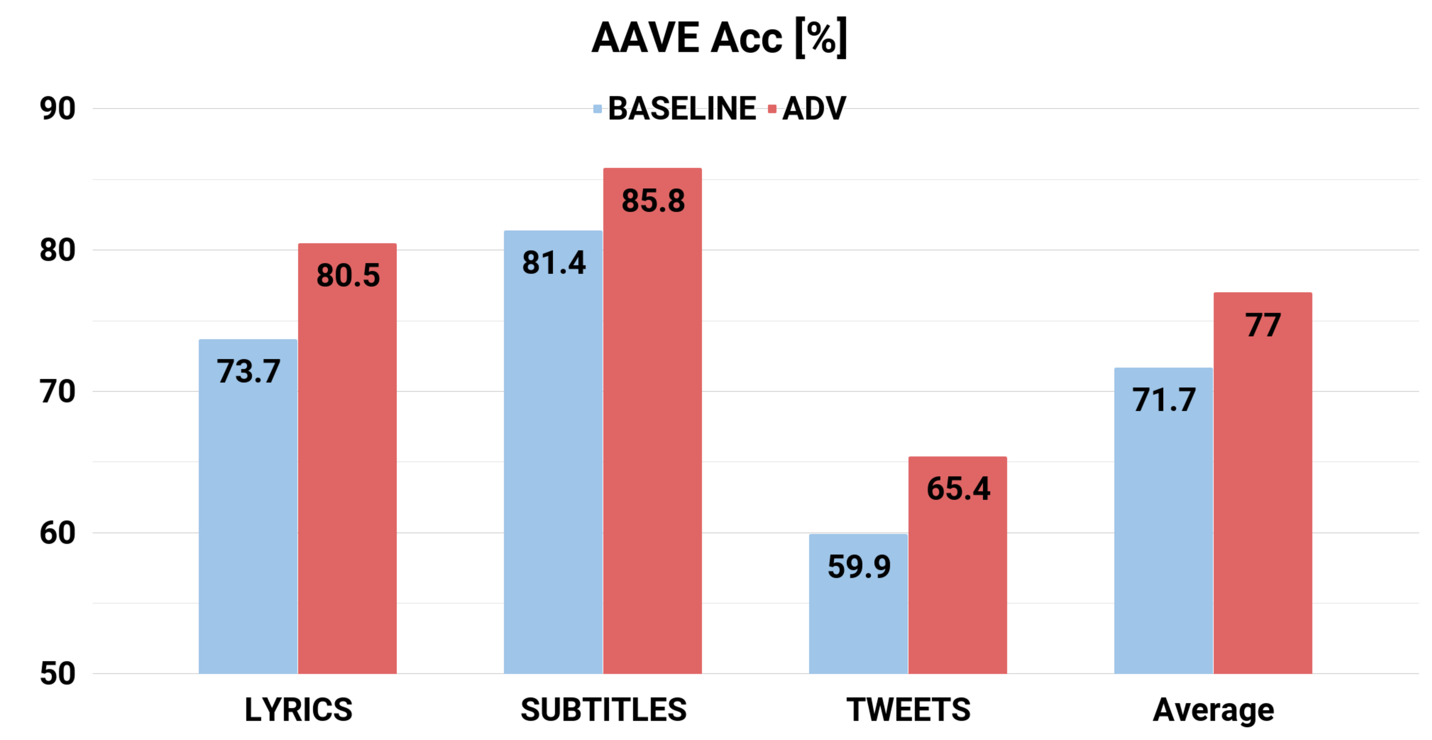
- experiment with two attributes:
  - GENDER: female (F) and male (M)
  - AGE: over-45 (O45) and under-35 (U35)





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

- Three heterogeneous domains: LYRICS, SUBTITLES and TWEETS



- Confusion Matrix: