Adversarial Removal of Demographic Attributes from Text Data

EMNLP 2018

Elazar and Goldberg, 2018

Gangwoo Kim

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

0 Introduction

We would like decisions to take into account factors which we deem to be irrelevant to decision

→ such as the gender, age and race of individual

Protected Attributes

irrelevant factors for constructing the predictive model

"textual information can be predictive of some demographic factors"

1 Learning Setup

task labels

a protected attribute

classifier

attacker

a representation

document

1 Learning Setup

Goal : We want decision y = f(x) to be
oblivious to z

Attacker
After the classifier c(h(x)) is fully trained,
we use the encoder h → z

Definition
A protected attribute has *leaked*
if we can train a classifier *h* → *z*

A protected attribute has *guarded*
if we cannot train it

2 Data, Tasks and Protected Attr.

Corpus - Twitter messages

1. **DIAL**
   - Task : binary emoji-based **sentiment** and binary **tweet-mention** prediction
     Sentiment : Positive v Negative
     Mention : conversational v non-conversational
   - Protected : race of authors
     Race : Light(Standard American English) v Dark (Else)

2. **PAN 16**
   - Task : to classify a given **dependency relation** between two tokens
   - Protected : Age and gender
     Age : (18-34) v (35+)
     gender : male v female

We collected 160K for training and 10K for development from each.
Each split is balanced with respect to both the main and the protected labels.

4 Data Leakage

Data	Task	Accuracy
DIAL	Sentiment	67.4
	Mention	81.2
	Race	83.9
PAN16	Mention	77.5
	Gender	67.7
	Age	64.8

Table 1: Accuracies when training directly towards a single task.

			Balanced		Unbalanced	
Data	Task	Protected Attribute	Task Acc	Leakage	Task Acc	Leakage
DIAL	Sentiment	Race	67.4	64.5	79.5	73.5
	Mention	Race	81.2	71.5	86.0	73.8
PAN16	Mention	Gender	77.5	60.1	76.8	64.0
		Age	74.7	59.4	77.5	59.7

Table 2: Protected attribute leakage: balanced & unbalanced data splits.

Left

- We begin by examining how well can we perform on each task

Right

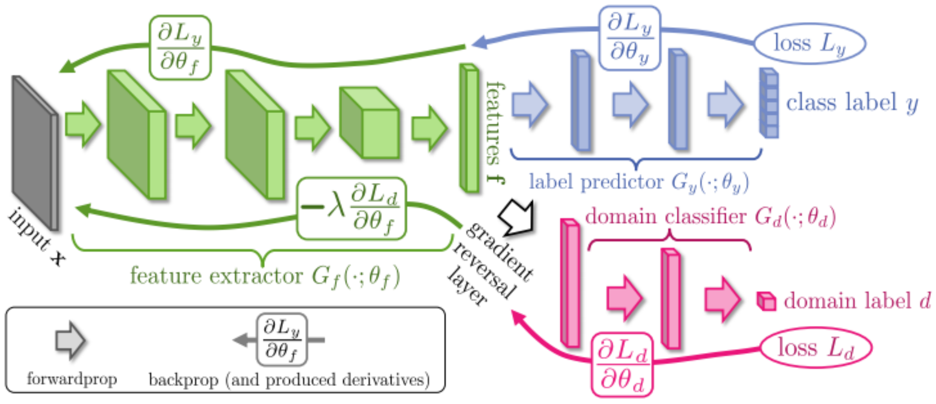
- … train the attacker network to predict the protected attributes based on a hidden rep.

5 Mitigating Data Leakage

Adversarial Training

- Objective : to create the rep h
  s.t. it's maximally informative for the main task,
  while at the same time minimally informative of the protected attribute

$$\arg\min_{h,c,adv} L(c(h(x_i)), y_i) + L(adv(g_\lambda(h(x_i))), z_i)$$



Data	Task	Protected Attribute	Task Acc	Leakage	Δ
DIAL	Sentiment	Race	64.7	56.0	5.0
	Mention	Race	81.5	63.1	9.2
PAN16	Mention	Gender	75.6	58.5	8.0
	Mention	Age	72.5	57.3	6.9

Table 3: Performances on different datasets with an adversarial training. Δ is the difference between the attacker score and the corresponding adversary's accuracy.

5 Mitigating Data Leakage

Strengthening the Adversarial Component

- **Capacity** of Adeversarial component – attacker's hidden dimension
- **Weight** – *tuning* λ
- **Ensemble** – of attacker

		DIAL			PAN16					
Method	Parameter	Sentiment	Race	Δ	Mention	Gender	Δ	Mention	Age	Δ
No Adversary Baseline	-	67.4	14.5	-	77.5	10.1	-	74.7	9.4	-
Standard Adversary	(300/1.0/1)	64.7	6.0	5.0	75.6	8.5	8.0	72.5	7.3	6.9
Adv-Capacity	500	64.1	6.7	5.2	73.8	8.1	6.7	71.4	**4.3**	4.1
	1000	63.4	7.1	4.9	75.2	8.9	7.0	71.6	6.3	4.0
	2000	65.2	8.1	6.9	76.1	6.7	6.4	71.9	6.0	5.7
	5000	63.9	**6.2**	3.7	74.5	5.6	1.6	73.0	10.2	9.6
	8000	65.0	7.1	4.8	75.7	**5.4**	4.2	71.9	9.8	7.3
λ	0.5	63.9	**6.8**	6.2	75.6	7.8	6.8	73.1	4.8	3.4
	1.5	64.9	7.4	5.4	75.6	**4.9**	2.4	72.5	6.8	5.8
	2.0	64.2	7.3	5.9	76.0	-7.2	6.7	72.1	8.5	7.7
	3.0	65.8	10.2	10.1	73.7	6.4	6.1	72.5	-6.3	5.2
	5.0	50.0	-	-	73.6	6.5	5.7	69.0	**3.2**	2.9
Ensemble	2	62.4	7.4	5.4	74.8	6.4	5.0	72.8	8.8	8.3
	3	66.5	6.5	5.0	75.3	4.9	3.1	72.1	6.7	6.0
	5	63.8	**4.8**	2.6	74.3	**4.1**	3.0	70.1	**5.7**	5.4

Table 4: Results of different adversarial configurations. **Sentiment/Mention**: main task accuracy. **Race/Gender/Age**: protected attribute recovery difference from 50% rate by the attacker (values below 50% are as informative as those above it). **Δ**: the difference between the attacker score and the corresponding adversary's accuracy. The bold numbers are the best *oblivious* classifiers within each configuration.

6 Analysis

Leakage via Embeddings

- Which part of encoder contributes to leakage

		Embedding	
		Leaky	Guarded
RNN	Leaky	64.5	67.8
	Guarded	59.3	54.8

Table 6: Accuracies of the protected attribute with different encoders.