brcsomnath/adversarial-scrubber

Adversarial Scrubbing of Demographic Information for Text Classification

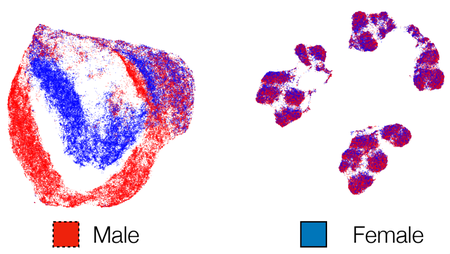
Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier B. Oliva, Shashank Srivastava, Snigdha Chaturvedi
{somnath, sayghosh, yiyuanli, joliva, ssrivastava, snigdha}@cs.unc.edu

UNC NLP | UNC DEPARTMENT OF COMPUTER SCIENCE

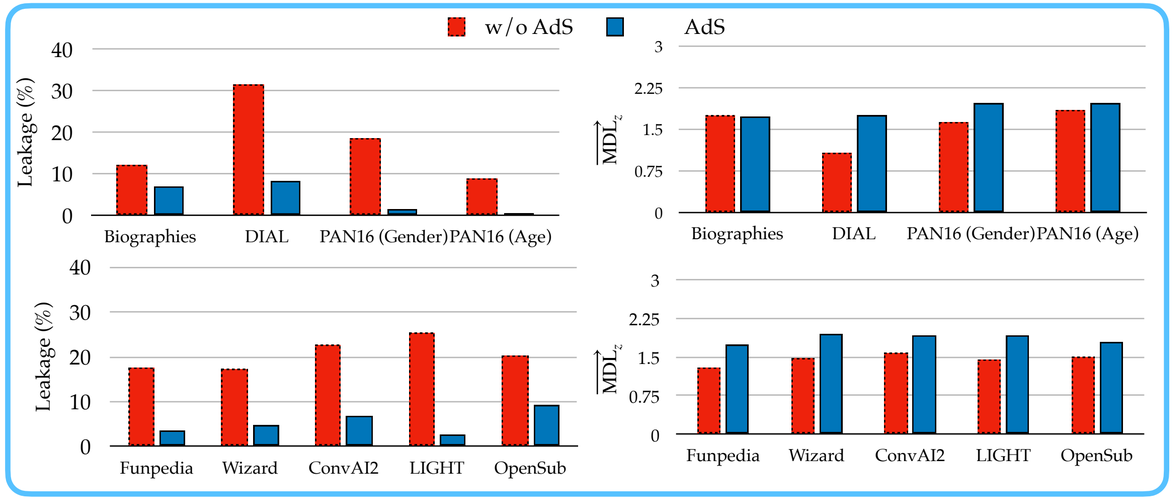
Problem Statement

- Text classification systems can be **biased** towards certain demographic groups
- Natural language is highly indicative of demographic attributes
- ML model representation **encode sensitive information** even without having direct access to them
- We propose AdS to learn **fair representations** during training text classification systems

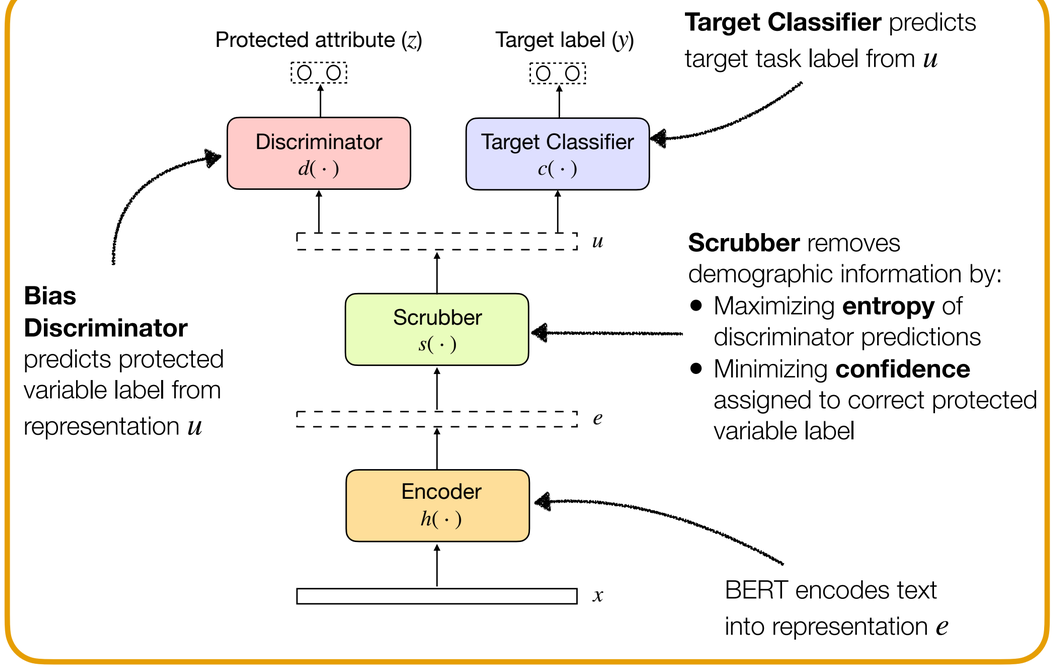
Visualization



Performance



Adversarial Scrubber



Target Classifier predicts target task label from u

Scrubber removes demographic information by:
- Maximizing **entropy** of discriminator predictions
- Minimizing **confidence** assigned to correct protected variable label

Bias Discriminator predicts protected variable label from representation u

BERT encodes text into representation e

Theoretical Results

- Minimizing Scubber loss increases the Bias discriminator loss
- Discriminator loss doesn't decrease when Encoder, Scrubber and Bias discriminator are updated
- For a strong Bias discriminator, the encoder and scrubber converge on the target task w/ o leaking demographic information

Probing

We evaluate representation u using off-the shelf sklearn classifier and evaluate the metrics below:
- Protected variable **accuracy** (lower value is desirable)
- Minimum description length (**MDL**) of protected variable (high value desired)
- Target variable accuracy (high value is desirable)

Conclusion

- We propose **Adversarial Scrubber** (AdS) to learn fair representation during text classification
- Theoretical and empirical analysis show AdS converges on target task **without leaking** sensitive information
- Effectively evaluate information content using MDL