Detecting and Reducing Bias in a High Stakes Domain

CS@CU Computer Science

[1]Ruiqi Zhong, [1]Yanda Chen, [2]Desmond Patton, [2]Charlotte Selous, [1]Kathleen McKeown
[1]Department of Computer Science, [2]School of Social Work, Columbia University

SAFE LAB

Topic Introduction

In cities such as Chicago, gang-involved youth increasingly turn to social media to post about emotions of ***loss*** when friends or family members are shot and killed. As grief turns to anger, their posts turn to ***aggression*** and ultimately into plans for revenge. Community outreach workers currently manually scour online spaces to identify such possibilities and intervene to diffuse situations. To scale their work and save critical time, we built an **automatic tool to identify Aggression or Loss in social media posts.** CS researchers developed machine learning algorithms to **classify tweets as "Aggression", "Loss" and "Other".**

This is a high stakes domain and machine learning systems are known to incorporate bias. For example, the COMPAS Recidivism Algorithm, which assesses the likelihood of a criminal defendant to re-offend and influences judges' sentencing decisions, is known to be biased against African Americans (Feller et al., 2016). To avoid similar errors, we aim to develop a system that fulfills the following criteria:

- Interpretable: model provides explanations to outreach workers and researchers.
- Fair: model uses legitimate features rather than dubious correlations to make predictions.
- Robust: model makes similar predictions when input is modified in a "non-perceptible" way.
- Performance: model maintains reasonable accuracy.

The problems are far from solved yet, but many of the elements appearing in this project will be relevant for similar/future works.

LOSS

Damn juss peeped shorty on tha news out here @[redacted] ..smh.. crazyy.. #RIPShorty

AGG.

My bro [redacted] thirsty he jus wana 👏 sum 💩 🔪 😈 💯

Data & Model

- 4,936 labeled tweets
  - Aggression: 329 tweets (6.7%), interannotator agreement = .94
  - Loss: 734 tweets (14.9%), interannotator agreement = .83
  - Other: 3,873 (78.4%)
- 1 million unlabeled tweets (snowball sampling -> 279 users)
- Using Context and Domain specific word embedding + Convolutional Neural Network we achieved 68% macro-F score.

Interpretability and Identifying Biases

- Used Locally Interpretable Model-agnostic Explanation (LIME) to interpret the model and identify the most influential unigram.
- Checked the second order consistency.
- **~10% times** the model considers the token "a"/"on" to be the top first or second most influential token.

Tweet	Score	Masked	influence
i a kill u ova mine	0.273	(original)	0
UNKNOWN a kill u ova mine	0.315	i	-0.042
i UNKNOWN kill u ova mine	0.263	a	0.009
i a UNKNOWN u ova mine	0.065	kill	0.207
...	...	...	...

Table 1: LIME, "kill" being the top influence word (having rank 1)

@user get up on me with you already know 👌 finish doing yo thang
grew up in this shit want let it hang on me
2 up cant down on me
@user aw i hear u tell em demo with me a lot more
made a man bitch n*** ima don 💯
can't let a mf get up on me

Confirming the Bias Through Adversary

- If we find a potential bias, the best way to confirm it is to break the system with this bias -> **build adversary**
- Appending "a" to every tweet **decreases precision by half** -> resulting tweets are unnatural.
- Optimizing and Confirming Adversary Naturalness:
  - 1) for a given tweet in the labeled set, add "a"/"on"/<other stop words> at each position.
  - 2) pick the most likely one scored by a language model trained on the unlabeled corpus.
  - 3) keep the top 800 edits in the labeled set as the "adversary"
  - 4) how many labels did it flip?
  - 5) ask domain experts to classify which tweets have been edited? -> accuracy = 75% (out of 36)

Systematic Debiasing & Rationale Annotation

- The domain experts annotate the most influential words.
  - 1) for evaluation: human & machine rationale should be similar.
  - 2) for training: reducing the effect of dubious correlations.

Tweets
"@user_ do you think it's cool yo ass slow ? You sound dumb asf"
"Rob you fucking goofys that got some to say about rage just to show y'all bitches"
"GlizzyGang Bitch We Got Out Glocks Up😈"

Table 1: Typical aggression tweets in our dataset, which includes annotations with human rationales (underlined).

Rationale Metrics and Training

- We developed the rationale rank (RR) metrics:
  - For a given tweet, rank tokens by influence score;
  - Define RR as the rank of expert rationale word.
  - Average across true positive predictions.
- Train attention on rationale words.

$$(h_1, h_2, ..., h_l) = LSTM(w_1, w_2, ..., w_l)$$

$$\alpha_i = tanh(\mathbf{v}h_i), A(i) = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}}$$

$$z = [\sum_{i=1}^{l} A(i)h_i, c]$$

$R = \{r_1, r_2 \dots r_d\}$ are indexes of the tokens that are rationale words.

$$A^*(i) = \frac{1}{d} \text{ if } i \in R, 0 \text{ otherwise}$$

$$\mathcal{L}_{attn} = KL(A^*||A), \mathcal{L} = \mathcal{L}_{clf} + \lambda_{attn}\mathcal{L}_{attn}$$

Unigrams / Models	a	on	da	into	of	that
Blevins et al. (2016)	44.16[!]	121.00[!]	37.44[!]	117.72[!]	36.84[!]	74.48 [!]
Chang et al. (2018)	26.56	36.00	8.52*	21.68	19.20	7.52*
CNN + Twitter	38.40	35.80	17.60	15.88	5.24*	7.64
LSTM	16.16	40.28	21.52	26.92	13.20	13.12
LSTM + Rationale	13.52*	23.16*	12.40	10.28*	11.40	8.96

Model	Avg RR	RR = 0	RR = 1
Blevins et al. (2016)	1.70	0.60	0.13
Chang et al. (2018)	1.42	0.54	0.17
CNN + Twitter	1.82	0.43	0.25
LSTM	1.73	0.50	0.15
LSTM + Rationale	**0.86**	**0.69**	0.14

An RR of 0 is the fraction of positive predictions in which the model's most influential word is a human rationale.