Contextualizing Hate Speech Classifiers with Post-hoc Explanations

Brendan Kennedy*, Xisen Jin*, Aida Mostafazadeh Davani, Morteza Dehghani, Xiang Ren
University of Southern California





ACL 2020 (Virtual) Short Paper Presentation

Paper: https://arxiv.org/abs/2005.02439
Code: https://github.com/BrendanKennedy/contextualizing-hate-speech-models-with-explanations

Bias in Hate Speech Data

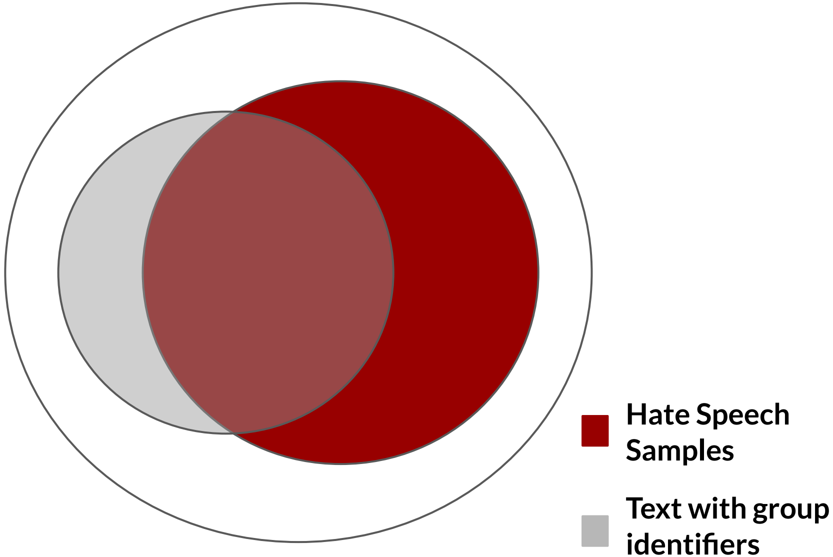
Group identifiers/social group terms are disproportionately frequent in hate speech data
Wiegand, Ruppenhofer & Kleinbauer (2019)

"There is a great discrepancy between whites and blacks in SA. It is … [because] blacks will always be the most backward race in the world."

But occur in many non-hate contexts as well:

"[F]or many Africans, the most threatening kind of ethnic hatred is black against black."

Training Data

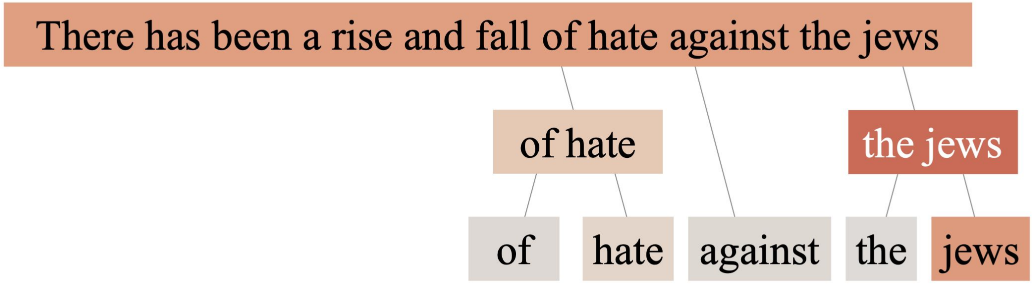


Problem Statement - Hate speech models treat the presence of group identifiers as an indicator of hate speech. But what matters is the group identifier *plus context*

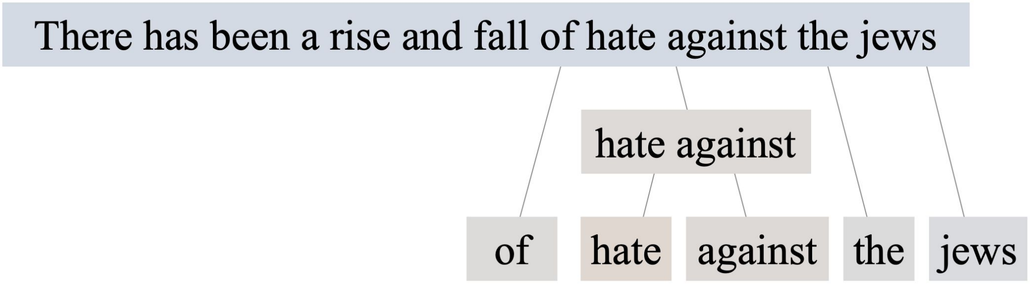
Understanding and Correcting Model Bias

We applied a **post-hoc explanation algorithm** (model agnostic) to quantify if models' predictions were biased towards group identifiers.

We found that false positive errors were caused by models **associate group identifiers with "hate"**



Our goal: neutralizing influence of group identifiers for non-hate contexts without performance loss on hate detection



Regularizing Post Hoc Explanations

Notations: w - group identifier words; **x** - input sentence; **s(·)**- model output; ***φ***(·) - explanation score,
S - set of all group identifiers; **L** - loss function

Step 1. Sampling-and-OCclusion (SOC) Explanations *(Jin et al., 2020)*

$$\phi(\mathbf{w}) = E_{\mathbf{x}_\delta}[\underbrace{s(\mathbf{x}) - s(\mathbf{x}\backslash\mathbf{w})}_{\text{Prediction difference when word w is masked}}]$$

Prediction difference when word w is masked

marginalized over contexts of word w around a fix-sized window x_δ

***φ*(*w*):** "How does the group identifier alone affect the prediction"?

Step 2. Regularizing Explanations of Group Identifier Terms

$$\mathcal{L} = \mathcal{L}' + \alpha \underbrace{\sum_{w \in x \cap S} [\phi(w)]^2}_{\text{penalizing explanation scores on group identifiers}}$$

penalizing explanation scores on group identifiers

Discourage making predictions with group identifier terms alone

Results of Regularization: Performance

Datasets

- Gab Hate Corpus (**GHC**; Kennedy et al., 2020)
- Stormfront (de Gibert et al. 2018)
- **NYT** (News articles, non-hate stratified sample across group identifiers)

Methods

- Vanilla BERT
- Identifiers removed before training (WR)
- Regularizing Input Occlusion explanations
- Regularizing SOC explanations (ours)

Training set	**GHC**			
Method / Metrics	**Precision**	**Recall**	**F1**	**NYT Acc.**
BoW	62.80	56.72	59.60	75.61
BERT	69.87 ± 1.7	66.83 ± 7.0	67.91 ± 3.1	77.79 ± 4.8
BoW + WR	54.65	52.15	53.37	89.72
BERT + WR	67.61 ± 2.8	60.08 ± 6.6	63.44 ± 3.1	89.78 ± 3.8
BERT + OC (α=0.1)	60.56 ± 1.8	**69.72** ± **3.6**	64.14 ± 3.2	89.43 ± 4.3
BERT + SOC (α=0.1)	**70.17** ± **2.5**	69.03 ± 3.0	**69.52** ± **1.3**	83.16 ± 5.0
BERT + SOC (α=1.0)	64.29 ± 3.1	69.41 ± 3.8	66.67 ± 2.5	**90.06** ± **2.6**

Results of Regularization: Term Importance

- Top 20 terms in each model (Vanilla BERT vs. SOC regularized BERT) by average SOC importance
- Change in rank importance (Δ Rank) between models
- Group identifiers highlighted

BERT	Δ Rank	Reg.	Δ Rank
ni**er	+0	ni**er	+0
ni**ers	-7	fag	+35
kike	-90	traitor	+38
mosques	-260	faggot	+5
ni**a	-269	bastard	+814
jews	-773	blamed	+294
kikes	-190	alive	+1013
nihon	-515	prostitute	+56
faggot	+5	ni**ers	-7
nip	-314	undermine	+442
islam	-882	punished	+491
homosexuality	-1368	infection	+2556
nuke	-129	accusing	+2408
niro	-734	jaggot	+8
muhammad	-635	poisoned	+357
faggots	-128	shitskin	+62
nitrous	-597	ought	+229
mexican	-51	rotting	+358
negro	-346	stayed	+5606
muslim	-1855	destroys	+1448

Conclusion and Future Work

Conclusion

- Bias can be addressed through *model enhancement* rather than *data augmentation*, by advancing explainability and developing techniques that operate on explanation algorithms like SOC

Unexplored Angles

- Our list of terms was *ad hoc*; lists provided by Dixon et al., 2018 can be applied
- Formal application of our approach to address fairness *between* social groups
- Explore other domains (e.g., Twitter), languages, and language models (e.g., GPT-2)

Paper: https://arxiv.org/abs/2005.02439
Code: https://github.com/BrendanKennedy/contextualizing-hate-speech-models-with-explanations