PEACE: Cross-Platform Hate Speech Detection - A Causality Guided Framework

Paras Sheth
Arizona State University, psheth5@asu.edu

Why do we need Generalizable Hate Speech Detection?

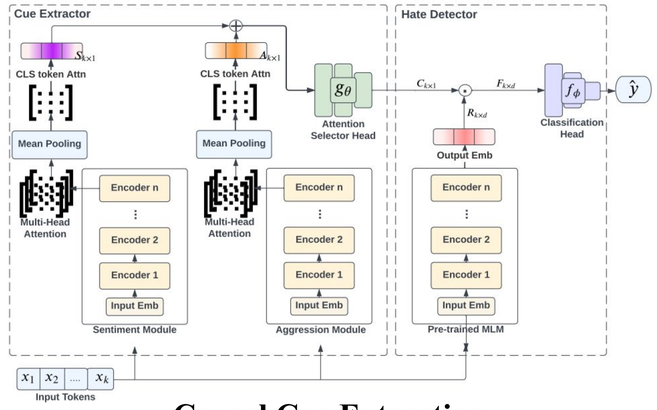
- Hate speech online entices violence, leading to **hate crimes** in real-world.
- Hate speech detection models need to be able to deal with the constant **growth and evolution** of hate speech.
- Current SOTA models were shown to be ineffective when dealing with **new targets** (e.g. Asian hate during COVID-19)

Current SOTA Methods - Limitations

- Linguistic cues based methods - suffer from shortcomings, such as linguistic methods form spurious correlations towards certain POS tags (e.g., adjectives and adverbs) or a particular category of words (e.g., abusive words)
- Auxiliary information based methods - are not extendable as the auxiliary information may not be available for large datasets or different platforms.

Proposed Method - PEACE

- When it comes to **explicit hate**, there are certain quantifiable properties that can aid in identifying hateful content - **aggression** and **sentiment**.



Causal Cue Extraction

- **Sentiment Module:** The sentiment module is a transformer encoder stack to predict the sentiment and the attention scores for tokens that contribute the most to the sentiment.
- **Aggression Module:** The aggression module is a transformer encoder stack to predict the aggression and the attention scores for tokens that contribute the most to the sentiment.
- We get the final attention vector $\mathbf{C_{k\ x\ 1}}$, by merging the attention weights from the two modules.

Hate Detector

- A transformer encoder stack to learn the semantic representation of the input F. The attention vector $C_{k\times1}$, is provided as an auxiliary signal.
- The hate detector predicts Y' by passing the product of F and $C_{k\times1}$ and passing it through a classification head.

Experiments

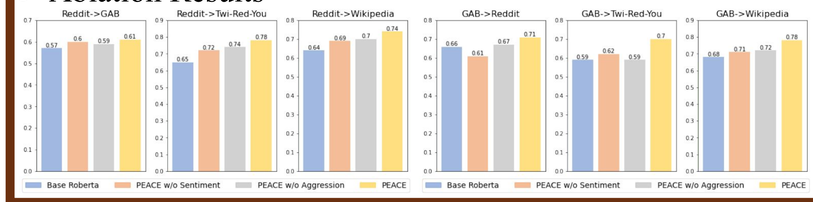
- Evaluating the generalizability across different platforms.

Source	Target	HateBERT	ImpCon (AugCon variant)	HateXplain	POS+EMO	PEACE
Twi-Red-You	GAB	0.58	0.58	0.60	0.54	**0.63**
	Reddit	0.71	0.64	**0.74**	0.54	**0.74**
	Wikipedia	0.71	0.70	0.70	0.60	**0.78**
	Twi-Red-You	**0.96**	0.94	0.92	0.87	0.95
	FRENK	0.46	0.44	0.48	0.45	**0.53**
GAB	GAB	**0.84**	0.65	**0.84**	0.76	0.76
	Reddit	0.69	0.64	0.70	0.56	**0.71**
	Wikipedia	0.74	0.64	0.70	0.49	**0.78**
	Twi-Red-You	0.61	**0.71**	0.61	0.59	0.70
	FRENK	**0.71**	0.57	0.60	0.59	0.69
Reddit	GAB	0.56	0.51	0.59	0.53	**0.61**
	Reddit	0.88	0.84	**0.89**	0.59	0.88
	Wikipedia	0.66	0.63	0.64	0.56	**0.74**
	Twi-Red-You	0.73	0.70	0.77	0.65	**0.78**
	FRENK	0.42	0.42	0.44	0.49	**0.54**
Wikipedia	GAB	0.65	0.63	0.64	0.56	**0.68**
	Reddit	0.73	0.71	**0.74**	0.58	0.72
	Wikipedia	0.95	0.93	0.86	0.94	**0.97**
	Twi-Red-You	0.73	0.72	0.74	0.69	**0.78**
	FRENK	0.60	0.51	0.61	0.52	**0.65**
FRENK	GAB	0.65	0.67	0.63	0.58	**0.69**
	Reddit	0.62	0.66	0.66	0.55	**0.71**
	Wikipedia	0.67	0.76	0.73	0.53	**0.81**
	Twi-Red-You	0.65	0.65	0.64	0.62	**0.78**
	FRENK	0.78	**0.79**	0.75	0.72	0.78

- Evaluating the generalizability across different targets.

Source	Target	HateBERT	ImpCon (AugCon variant)	HateXplain	POS+EMO	PEACE
Migrants	LGBTQ	0.74	0.68	0.65	0.61	**0.78**
LGBTQ	Migrants	0.66	0.67	0.64	0.58	**0.72**

- Ablation Results



References:
[1] Sheth, P., Kumarage, T., Moraffah, R., Chadha, A. and Liu, H., 2023. PEACE: Cross-Platform Hate Speech Detection-A Causality-guided Framework
Sponsored by the ONR grant # N00014-21-1-4002, the ARO garnet # W911NF2110030, and DARPA grant # HR001120C0123