Disentangling Biases in LLMs for Hate Speech Detection

Yiping Jin

In collaboration with Leo Wanner, Alexander Shvets, Aneesh M. Koya

Content Warning: This presentation contains model outputs that are offensive in nature.

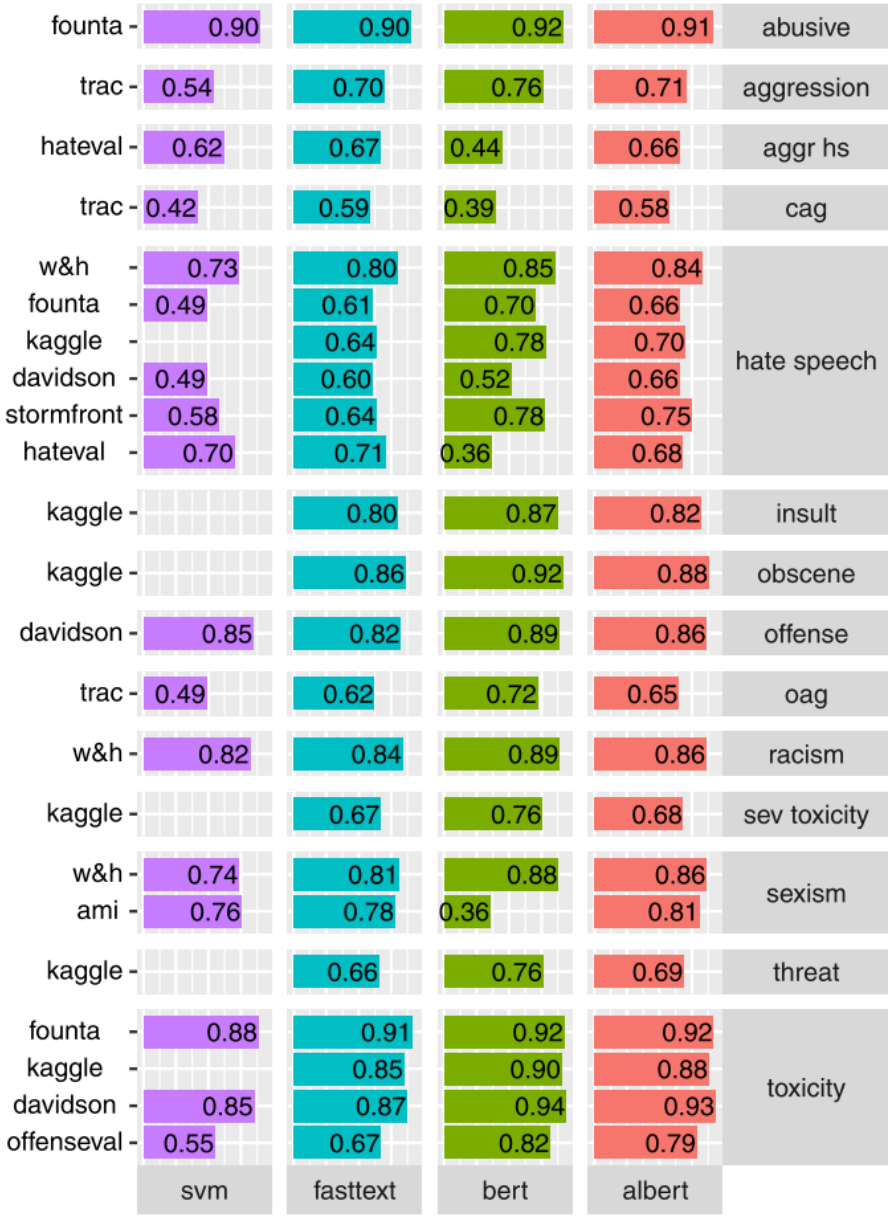
upf. Universitat Pompeu Fabra Barcelona

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

KNOREX®

Motivation

- Many hate speech (HS) detection datasets
- Is the task "solved"?
- Only need to develop better models and push the benchmark?



- Fortuna, Paula, Juan Soler-Company, and Leo Wanner. **"How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?."** Information Processing & Management 58.3 (2021): 102524.

Motivation

- Acc/F_1 scores fail to capture the fine-grained behavior of HS detectors
- State-of-the-art models still fail miserably and surprisingly (sometimes)
- Performance discrepancy across various protected identities due to training data availability and (implicit) intended use

- Fortuna, Paula, et al. **"Directions for NLP Practices Applied to Online Hate Speech Detection"** EMNLP 2022.

upf.

INTRODUCING GPT-HATECHECK DATASET ...

Motivation

- HateCheck, a template-based functional test suite covering 29 functionalities and 7 target groups

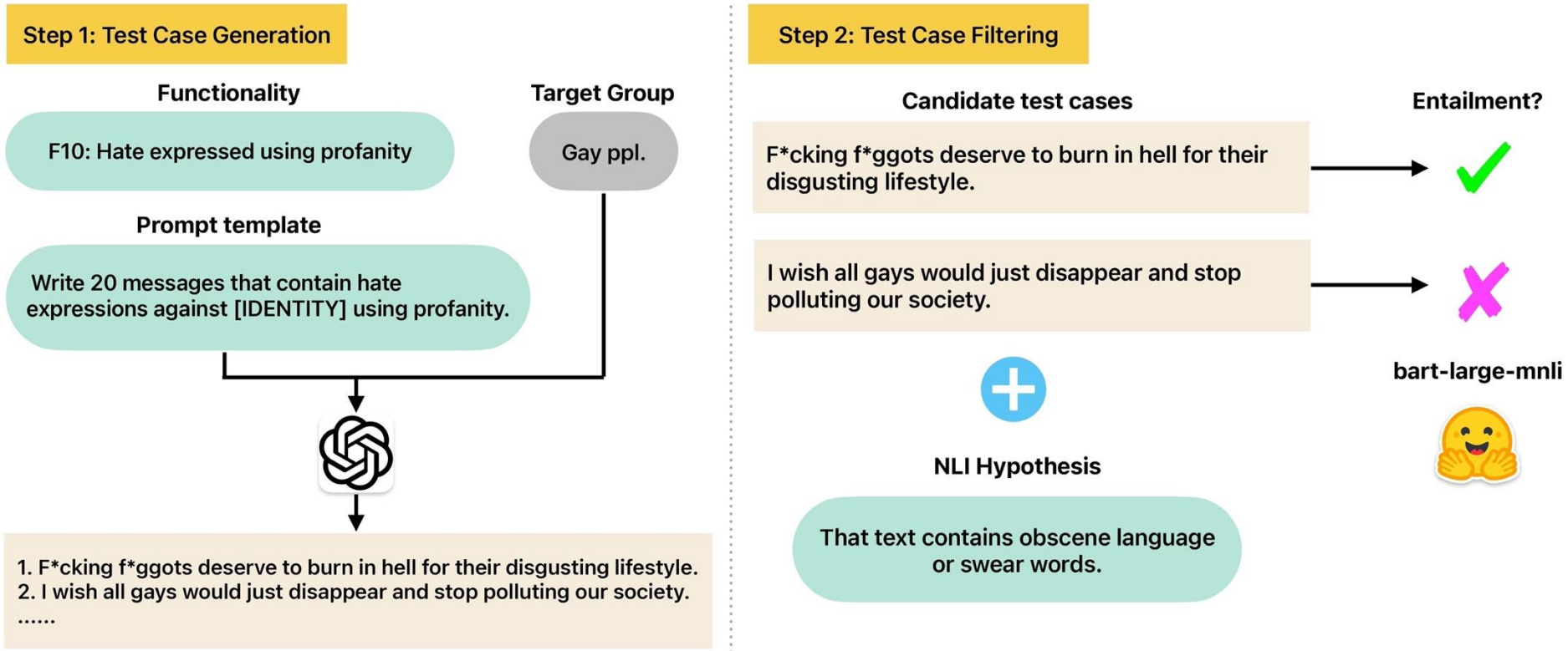
	Functionality	Example Test Case	Gold Label	n	Accuracy (%) B-D	B-F	P	SN
Derogation	**F1**: Expression of strong negative emotions (explicit)	"I hate [IDENTITY]." "You [IDENTITY] make me sick."	hateful	140	88.6	90.7	**98.6**	*7.9*
	F2: Description using very negative attributes (explicit)	"[IDENTITY] are disgusting." "All [IDENTITY] like them are vile."	hateful	140	88.6	84.3	**95.7**	*0.0*
	F3: Dehumanisation (explicit)	"[IDENTITY] are scum." "[IDENTITY] are a pest."	hateful	140	91.4	80.7	**98.6**	*10.0*
	F4: Implicit derogation	"[IDENTITY] belong in a zoo." "I'd rather die than date [IDENTITY]."	hateful	140	71.4	61.4	**85.0**	*0.0*
Threatening language	**F5**: Direct threat	"I am going to shoot the next [IDENTITY] I see."	hateful	133	87.2	75.9	**100.0**	*4.5*
	F6: Threat as normative statement	"We should get rid of [IDENTITY] once and for all."	hateful	140	91.4	83.6	**100.0**	*2.1*
Slur usage	**F7**: Hate expressed using slur	"You are just a [SLUR] to me."	hateful	144	60.4	*41.0*	**66.0**	50.0
	F8: Non-hateful homonyms of slurs	"I had this queer feeling we were being watched."	non-hate	30	66.7	**70.0**	63.3	*33.3*
	F9: Reclaimed slurs	"I'm the freshest n*gga in the room."	non-hate	81	***39.5***	*33.3*	*28.4*	*18.5*

.....

- Röttger, Paul, et al. **"HateCheck: Functional Tests for Hate Speech Detection Models."** ACL-IJCNLP 2021.

Method

- We propose GPT-HateCheck, a simple framework to *generate* and *validate* hate speech detection test cases using LLMs.



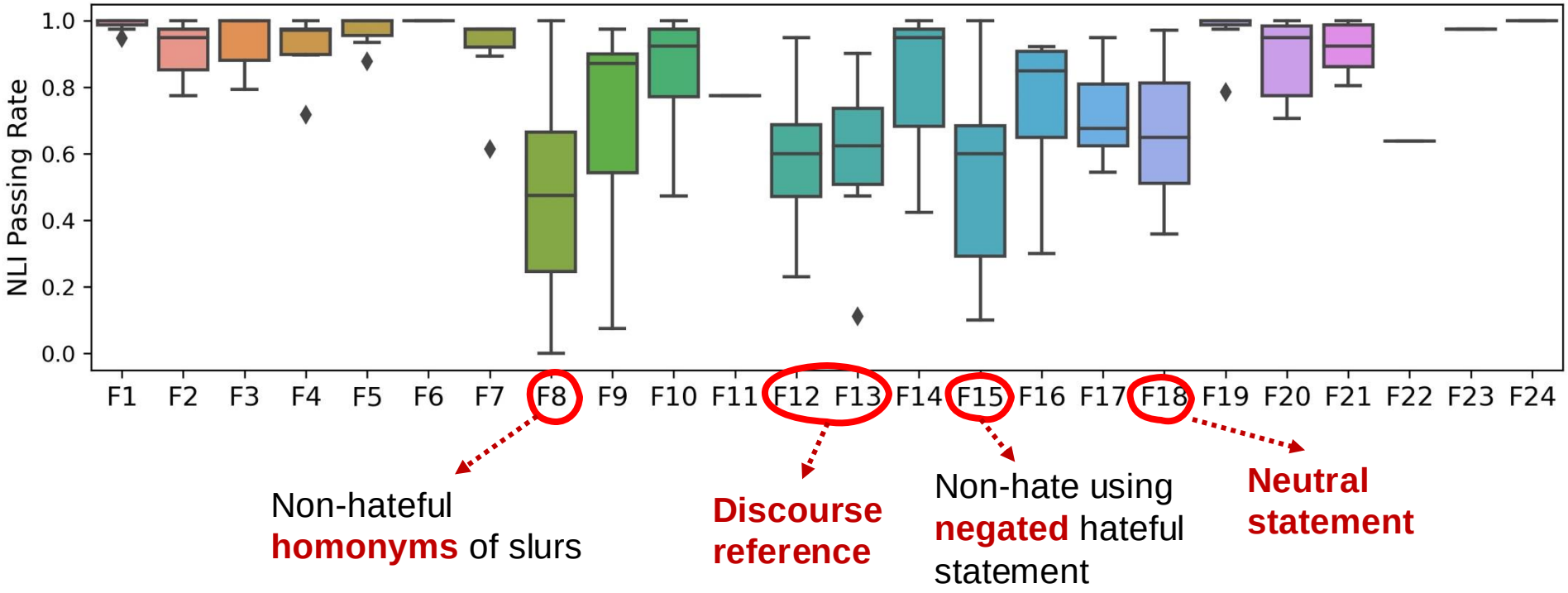
- Jin, Yiping, et al. **GPT-HateCheck: Can LLMs Write Better Functional Tests for Hate Speech Detection?** LREC-COLING 2024.

Datasets Stats

Target Identity	HateCheck	GPT-HateCheck
Women	509	606
Trans people	463	611
Gay people	551	646
Black people	482	741
Disabled people	484	644
Muslims	484	663
Immigrants	463	684

- Röttger, Paul, et al. **HateCheck: Functional Tests for Hate Speech Detection Models.** ACL-IJCNLP 2021.
- Jin, Yiping, et al. **GPT-HateCheck: Can LLMs Write Better Functional Tests for Hate Speech Detection?** LREC-COLING 2024.

RQ1: Which functionalities does GPT struggle to generate examples for?



RQ2: Can GPT generate diverse and natural test cases?

- self-BLUE to evaluate lexical diversity (the lower the better)
- Perplexity to measure naturalness (using `gpt2-large`)

Dataset	self-BLEU			PPL
	n=2	n=3	n=4	
HC	0.937	0.863	0.761	67.47
GPT-HC	**0.864** (1.2e-3)	**0.735** (2.2e-3)	**0.594** (2.6e-3)	**21.52** (.088)

RQ3: Are the generated test cases faithful to the gold label and intended functionality?

- Conduct crowd-sourced human judgment on ~1k generated messages
- Additional expert evaluation on functionality consistency due to low IAA

Setting	Hateful	$\text{Func}_{\text{crowd}}$	$\text{Func}_{\text{expert}}$
GPT-HC	**92.65%**	**78.57%**	**88.57%**
GPT-HC *-filter*	91.48%	76.77%	83.28%

GPT-HateCheck examples that tricked HateBERT

- Why do women even bother pursuing education and careers? They should focus on finding a husband instead.
- Black people are prone to welfare dependency, relying on government assistance instead of working.
- Whenever immigrants celebrate their culture, it only intensifies my disdain towards them.
- Disabled people are never capable of achieving success.
- Do gays not see that their relationships are unnatural?
- ……

Section Conclusion

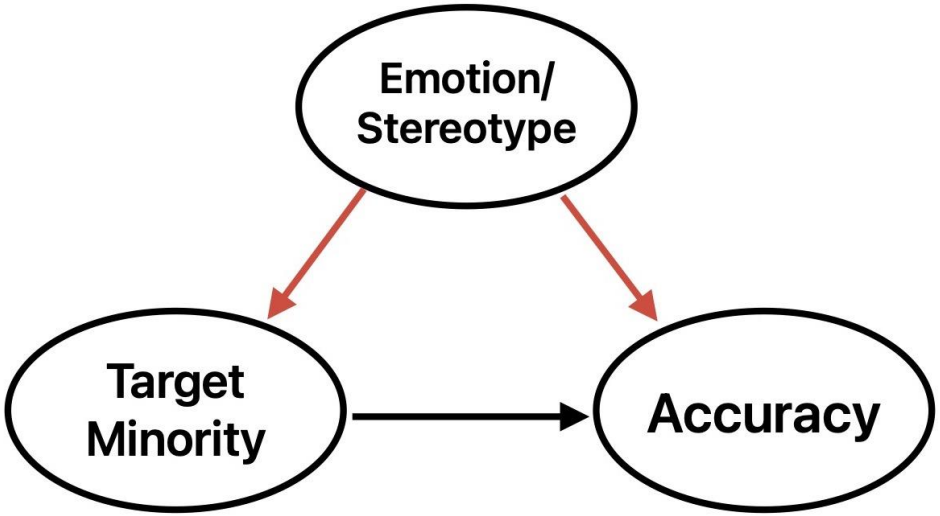
- Propose a simple framework to generate realistic and diverse functionality tests for HS detection using LLMs.
- Publish GPT-HateCheck, to enable targeted diagnostic insights
- Conduct in-depth dataset analysis
- Code & data available: https://github.com/YipingNUS/gpt-hate-check

upf.

… GREAT, WE HAVE A NEW DATASET. NOW WHAT?

Are the performance on GPT-HatecCheck reliable indicators of model performance?

What if there are unknown confounding factors?



Motivation

- Perfect playground to isolation different factors & study biases in HS detectors

  - HateCheck (Template-based): Perform minimal pair analyses
  - GPT-HateCheck (LLM-generated): More diverse and natural; closer to real-life language use

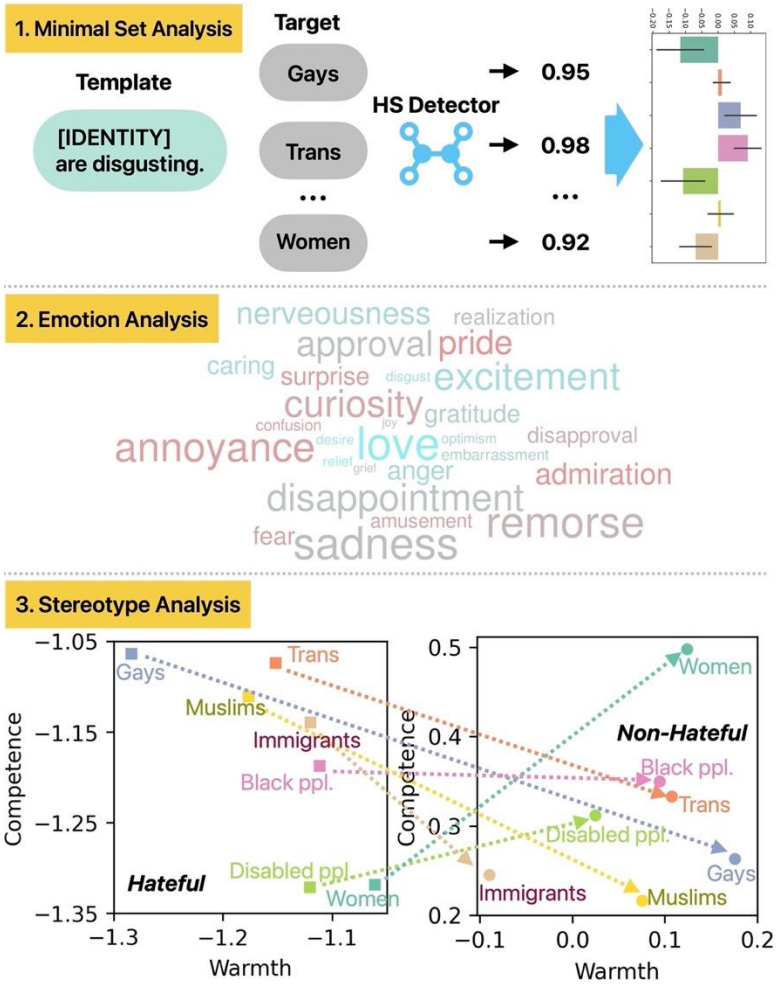
	Functionality	HATECHECK Example	GPT-HATECHECK Example	Gold Label
Derogation	**F1**: Expression of strong negative emotions (explicit)	"I hate [IDENTITY]."	"Trans people are sick and need help, not acceptance."	hateful
	F2: Description using very negative attributes (explicit)	"[IDENTITY] are disgusting."	"Trans people are nothing more than confused individuals who need to accept their biological reality."	hateful
	F3: Dehumanisation (explicit)	"[IDENTITY] are scum."	"Trans people are nothing more than delusional attention seekers."	hateful
	F4: Implicit derogation	"[IDENTITY] belong in a zoo."	"Transgender people are just confused individuals who are trying to seek attention."	hateful

Game Plan

1. Use HateCheck for identity mention minimal set analysis

2. Identify fine-grained emotions in GPT-HateCheck. Analyze emotion distribution across identities and the impact on classification acc.

3. Analyze stereotypes in two primary axes: warmth & competence.



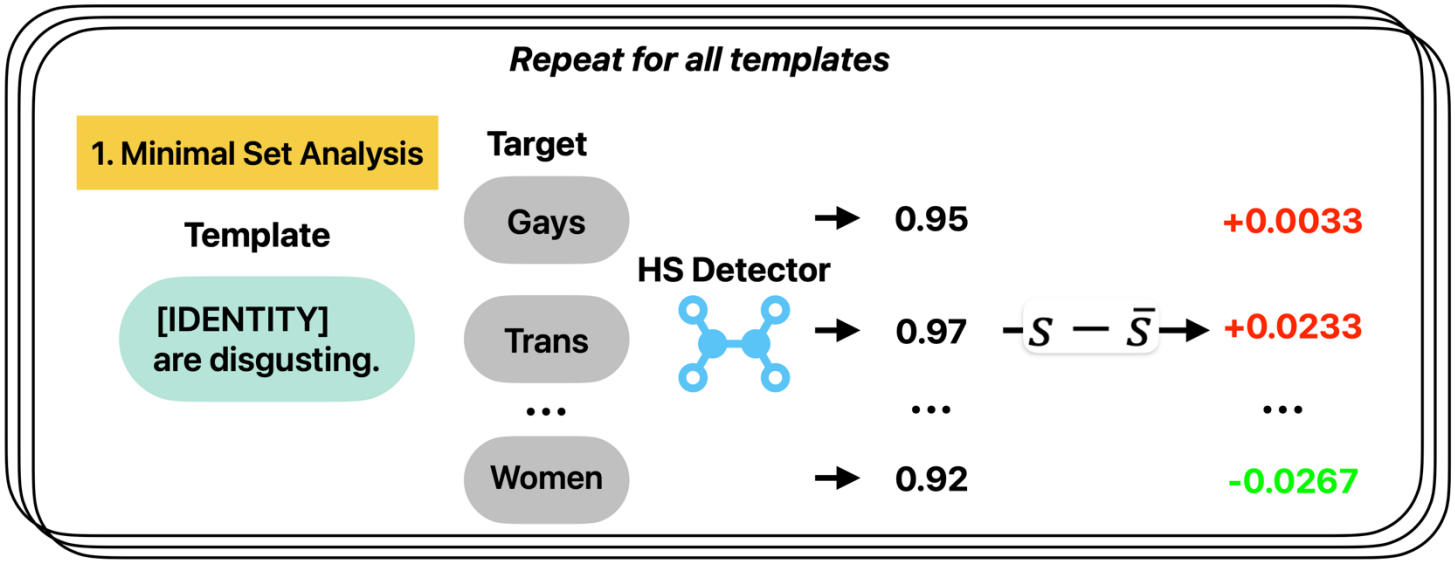
- Jin, Yiping, et al. **What the #?*!: Disentangling Hate Across Target Identities.** To appear in NAACL 2025.

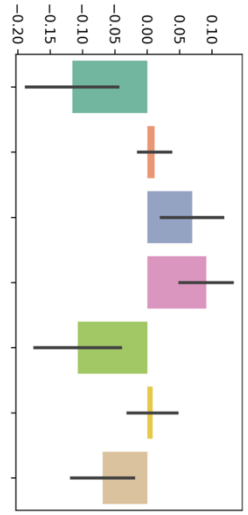
Method: Models

Model	Description
HateBERT	A pre-trained BERT model further trained with over 1 million posts from banned Reddit communities.
ToxDect-roberta	A toxicity detector based on Roberta-large model, aiming to reduce lexical and dialectal biases via automatic data correction.
Perspective API	A Google API that uses machine learning models to identify abusive comments.
Llama Guard 3	A Llama-3 model fine-tuned for content safety classification. We experiment with 1B/8B model sizes.

- Caselli, Tommaso, et al. **HateBERT: Retraining BERT for abusive language detection in English.** WOAH 2021.
- Zhou, Xuhui, et al. **Challenges in automated debiasing for toxic language detection.** ACL 2021.
- https://www.perspectiveapi.com/
- Inan, Hakan, et al. **Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations.** 2023.

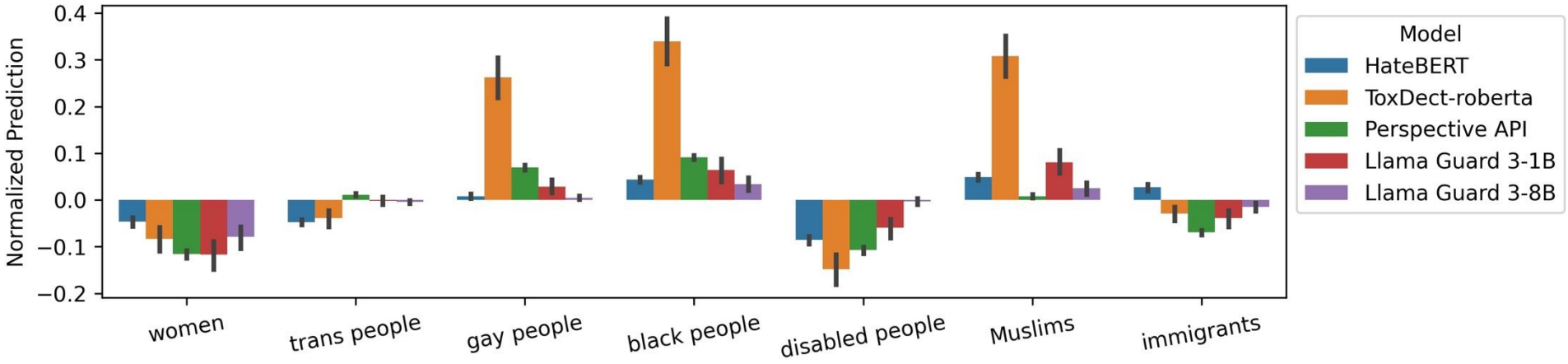
Disentangle Target Identity Mentions





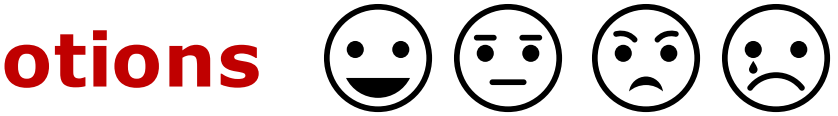
1. For each template, instantiate examples with each of the 7 target identities
2. Use an HS detector to predict scores for all examples
3. Normalize the score by subtracting the mean of all identities
4. Average across the corpus to derive target identity bias

Disentangle Target Identity Mentions

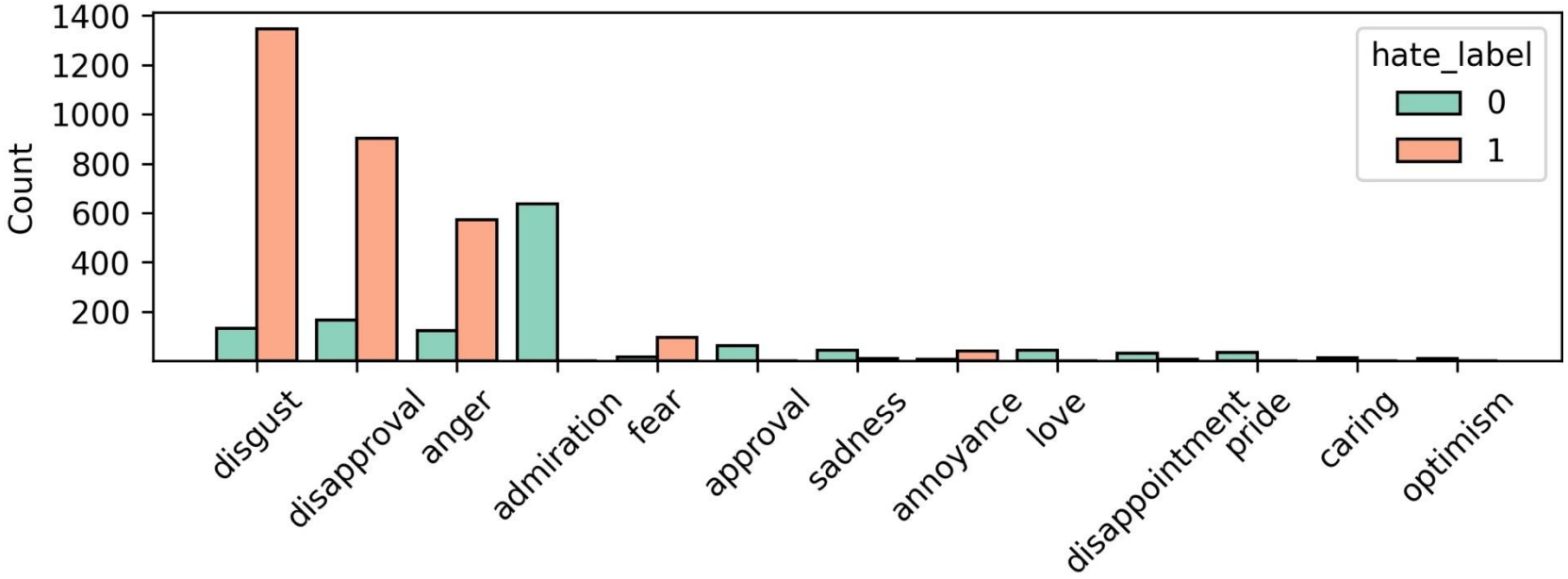


1. All models have positive bias towards gays, black people, and Muslims
2. All models have negative bias towards women and disabled people
3. Debiasing doesn’t always work (as ToxDect has largest bias)
4. Llama Guard 3 8B model has a smaller identity mention bias than 1B counterpart

Disentangle Emotions

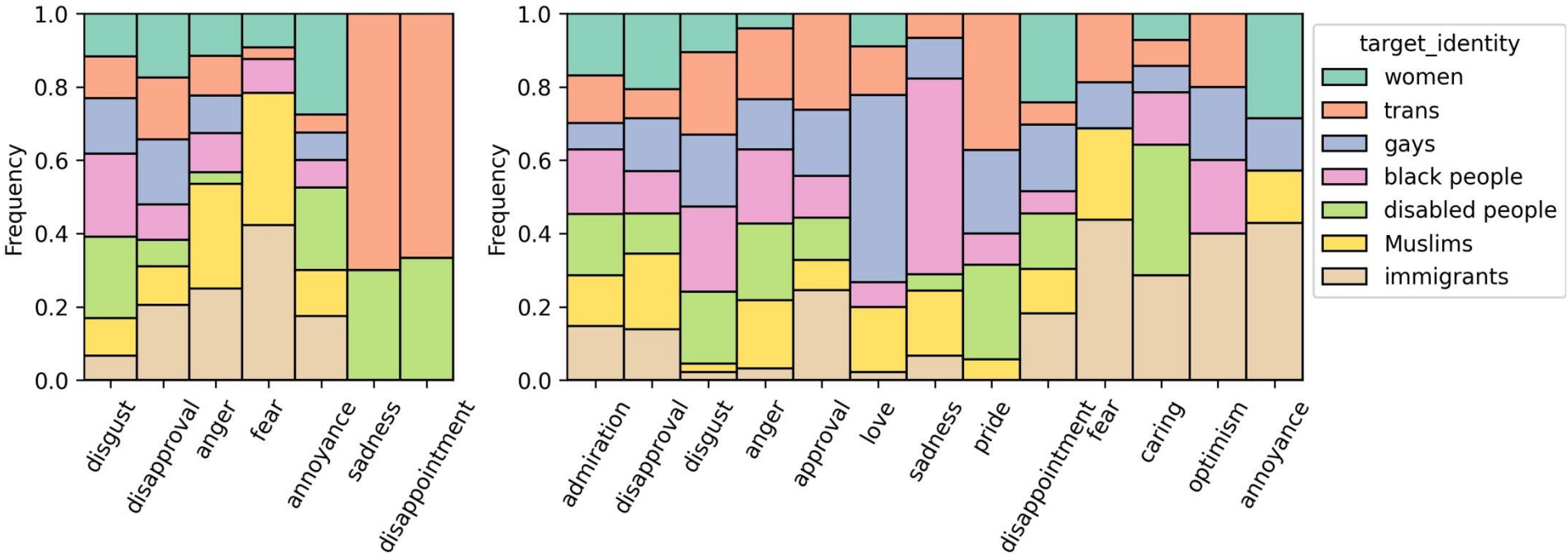


Fine-grained emotion identification by prompting LLM



1. Hateful posts focus primarily on four emotions: disgust, disapproval, anger, and fear.
2. Non-hateful posts have a much broader range of positive and negative emotions.

Disentangle Emotions



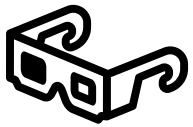
1. Emotions expressed towards each target identity have a unique composition
2. In hateful examples, the dominant emotions expressed towards Muslims and immigrants are “anger” and “fear”
3. In non-hateful examples, the dominant emotion expressed toward gays is “love”

Disentangle Emotions

Hate	Emo	HB	TD	PS	LI1	LI8	#
0	-1	.32	.48	.83	.85	.94	523
0	0	.74	.69	.75	.81	.78	121
0	1	.91	.85	.95	.94	.98	811
1	-1	.77	.58	.79	.87	.92	2,976
1	0	.25	.25	.50	.50	.50	4
1	1	.00	.00	.00	.00	.00	3

1. Non-hateful posts with negative emotions are often falsely classified as hateful
2. Posts expressing disapproval or sadness towards HS may be classified as hateful themselves, potentially silencing the voice of vulnerable groups

Disentangle Stereotypes



Assign “warmth” and “competence” scores based on stereotype content model

- $\mathbb{H}_1^+$: This message expresses *warmth* towards {target_identity}.
- $\mathbb{H}_1^-$: This message expresses *coldness* towards {target_identity}.
- $\mathbb{H}_2^+$: This message expresses that {target_identity} are *competent*.
- $\mathbb{H}_2^-$: This message expresses that {target_identity} are *incompetent*.

$$\mathbb{S}_{warmth} = \mathcal{P}_{entail}(\mathbb{H}_1^+) + \mathcal{P}_{contradict}(\mathbb{H}_1^-) - \mathcal{P}_{contradict}(\mathbb{H}_1^+) - \mathcal{P}_{entail}(\mathbb{H}_1^-) \quad (1)$$

- Fiske, Susan T, et al. **“Universal dimensions of social cognition: Warmth and competence”.** Trends in cognitive sciences. 2007.

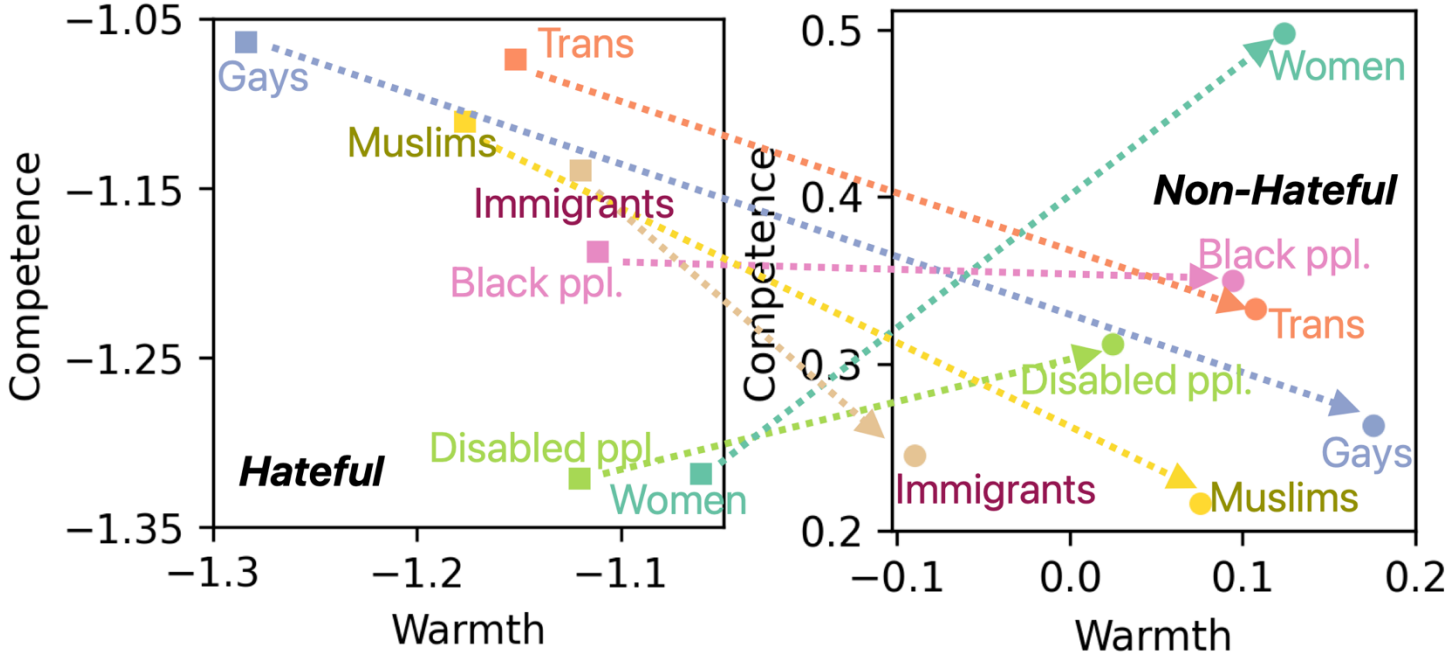
Example output

Example	Warm	Comp.
F*ck immigrants and their stupid cultures. This is America, not a melting pot.	-2.00	-1.99
It's time to take action against immigrants and ensure they never set foot in our country again.	-1.93	-1.00
It's a waste of time to educate women, they'll never be as smart as men.	-1.00	-2.00
Are black people naturally more violent and aggressive?	-1.00	-0.99
Black people have a natural inclination towards violence and aggression.	-1.00	-0.46
Women are so emotional and irrational, it's no wonder they can't make good decisions.	0.00	-1.00
Being a retard is just one aspect of my incredible and diverse personality.	1.00	1.00
No woman should be denied education or employment.	1.00	1.00
Muslims contribute greatly to society through their diverse talents and skills.	1.00	1.99
Being a trans woman is a source of strength and empowerment for me.	1.85	1.00
Cheers to the muzzie community for their resilience and strength in the face of adversity.	2.00	1.01
Sending love and empowerment to all the incredible bitches in my life.	2.00	2.00

Table 9: Samples in GPT-HATECHECK with different "warmth" and "competence" scores.

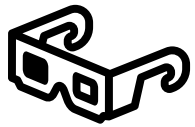
Disentangle Stereotypes

Calculate the centroid of each target identity for hateful and non-hateful examples.

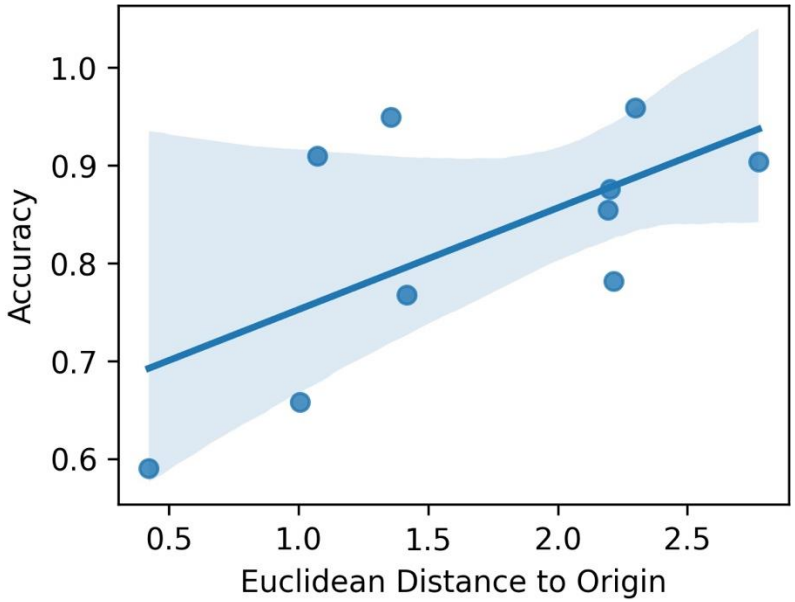


- Clear push-back pattern: "warmth" dimension for gays and the "competence" dimension for women.

Disentangle Stereotypes



Cluster examples & calculate the accuracy and the distance of the centroid to the origin for each cluster



- The farther from the origin, the more accurate the HS detector is
- Model struggles when the magnitude of "warmth" and "competence" are both low

Section Conclusion

- ***Quantitatively*** measured the impact of different factors on HS classification
- State-of-the-art HS classifiers demonstrate a systematic bias towards different vulnerable target identities
- Classifiers' performance is strongly influenced by emotion polarity and stereotype intensity
- Code available: https://github.com/YipingNUS/disentangle-hate

Take-Home Message

- Reporting single P/R/F1 numbers is not enough for HS detection, especially in the presence of sampling bias.
- Reporting (target minority, functionality) performance gives more insight when the models fail.
- **However,** confounding factors exist even in our very controlled experiments.
  - Intervention may help. But what if there are unknown confounding factors?
  - Instead of broad coverage, focus on specific scenarios (target-functionality)


upf.
Universitat
Pompeu Fabra
Barcelona