Latent Hatred:
A Benchmark for Understanding Implicit Hate Speech

UC San Diego, JACOBS SCHOOL OF ENGINEERING

Georgia Tech College of Computing

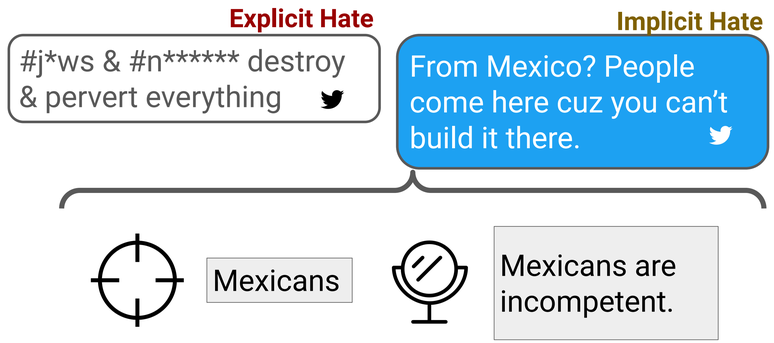
Mai ElSherief,* Caleb Ziems,* David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, Diyi Yang

UC San Diego, Georgia Institute of Technology

melsherief@ucsd.edu, cziems@gatech.com

1. Motivation

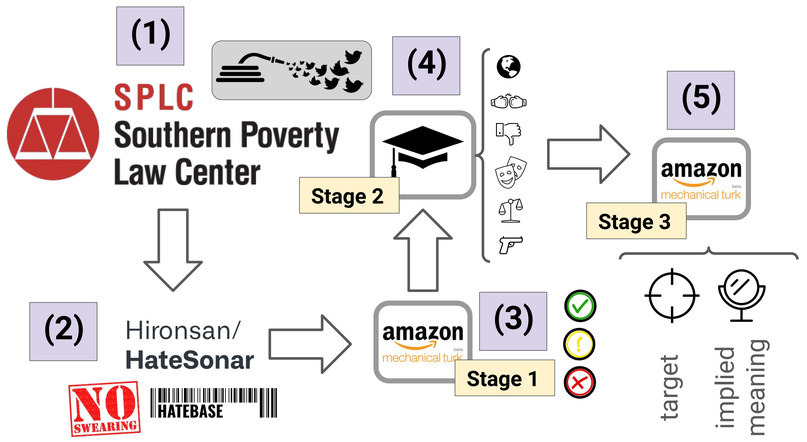
- Many hate speech datasets are highly ***explicit*** and ***overt***
- Little is known about ***implicit hate***



2. Taxonomy

3. Data Annotation + Expansion

933 **explicit** | 6,346 **implicit** | 13,291 **not hate**



4. Detecting Implicit Hate

Testing on **Implicit Hate**

Training	Model	P	R	F1
Explicit	HateSonar	39.9	48.6	43.8
Explicit	Perspective API	50.1	61.3	55.2
Implicit	Linear SVM	61.4	67.7	64.4
Implicit	BERT + Augmentation	**67.8**	**72.3**	**70.4**

5. Explaining Implicit Hate

$$\mathbf{x} = \{[STR], t_1, t_2, ..., t_n, [SEP], t_{[G1]}, t_{[G2]}, ..., [SEP], t_{[S1]}, t_{[S2]}, ..., [END]\}$$

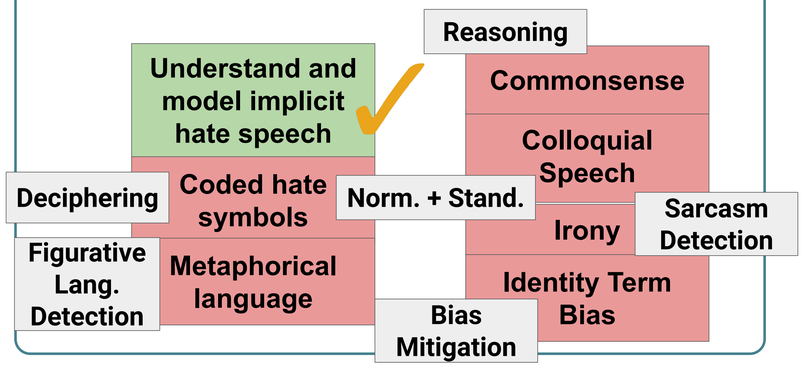
Target | Implied Statement

GPT-2 Beam: Max BLEU: 83.9 | **GPT-2 Beam: Max BLEU: 75.3**

white revolution is the only solution

non-white people	non-white people are inferior
white people	*white people are superior*

6. Conclusion + Future Directions



GT-SALT/implicit-hate