Latent Hatred: A Benchmark for Understanding Implicit Hate Speech

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

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Georgia Tech College of Computing

Introduction: A Benchmark for Implicit Hate Speech

⚠️ **Content Warning:** may contain upsetting examples.

Hate Speech Datasets

Basile et al. (2019)	Davidson et al. (2017)	Djuric et al. (2015)	Founta et al. (2018)
Burnap and Williams (2014)	Gao and Huang (2017)	Warner and Hirschberg (2012)	Waseem and Hovy (2016)
de Gibert et al. (2018)	Kennedy et al. (2018)	Sap et al. (2020)	Zampieri et al. (2019)

Hate Speech Datasets

Basile et al. (2019)	Davidson et al. (2017)	Djuric et al. (2015)	Founta et al. (2018)
Burnap and Williams (2014)	Gao and Huang (2017)	Warner and Hirschberg (2012)	Waseem and Hovy (2016)
de Gibert et al. (2018)	Kennedy et al. (2018)	Sap et al. (2020)	Zampieri et al. (2019)

Seeded w/ Hate Lexicon

Hate Speech Datasets Are:

☑ Explicit

● Basile et al. (2019)	● Davidson et al. (2017)	Djuric et al. (2015)	● Founta et al. (2018)
Burnap and Williams (2014)	Gao and Huang (2017)	Warner and Hirschberg (2012)	● Waseem and Hovy (2016)
de Gibert et al. (2018)	Kennedy et al. (2018)	Sap et al. (2020)	Zampieri et al. (2019)

● Seeded w/ Hate Lexicon

Hate Speech Datasets Are:

☑ Explicit

● Basile et al. (2019)	● Davidson et al. (2017)	Djuric et al. (2015)	● Founta et al. (2018)
Burnap and Williams (2014)	Gao and Huang (2017)	Warner and Hirschberg (2012)	● Waseem and Hovy (2016)
de Gibert et al. (2018)	Kennedy et al. (2018)	Sap et al. (2020)	Zampieri et al. (2019)

● Seeded w/ Hate Lexicon

% Expletives in Hate Class

10% 20% 30% 70%

Hate Speech Datasets Are:

- [x] Explicit
- [x] Targeted

Basile et al. (2019)	Davidson et al. (2017)	Djuric et al. (2015)	Founta et al. (2018)
Burnap and Williams (2014)	Gao and Huang (2017)	Warner and Hirschberg (2012)	Waseem and Hovy (2016)
de Gibert et al. (2018)	Kennedy et al. (2018)	Sap et al. (2020)	Zampieri et al. (2019)

Seeded w/ Hate Lexicon

Seeded w/ Racial Identifiers

% Expletives in Hate Class

10% 20% 30% 70%

Hate Speech Datasets Are:

- ☑ Explicit
- ☑ Targeted

Anti-Immigrant	Davidson et al. (2017)	Djuric et al. (2015)	Founta et al. (2018)
@ Lee Rigby Murder	Gao and Huang (2017)	**Anti-Semitism**	**Sexism, Racism**
de Gibert et al. (2018)	Kennedy et al. (2018)	**Stereotypes**	**Political Phrases**

- Seeded w/ Hate Lexicon
- Seeded w/ Racial Identifiers

% Expletives in Hate Class

10% 20% 30% 70%

Hate Speech Datasets Are:

- [x] Explicit
- [x] Targeted

Anti-Immigrant	Davidson et al. (2017)	Djuric et al. (2015)	Founta et al. (2018)
@ Lee Rigby Murder	Gao and Huang (2017)	**Anti-Semitism** (Warner and Hirschberg (2012))	**Sexism, Racism** (Waseem and Hovy (2016))
de Gibert et al. (2018)	Kennedy et al. (2018)	**Stereotypes**	**Political Phrases**

- Seeded w/ Hate Lexicon
- Seeded w/ Racial Identifiers

% Expletives in Hate Class

10% | 20% | 30% | 70%

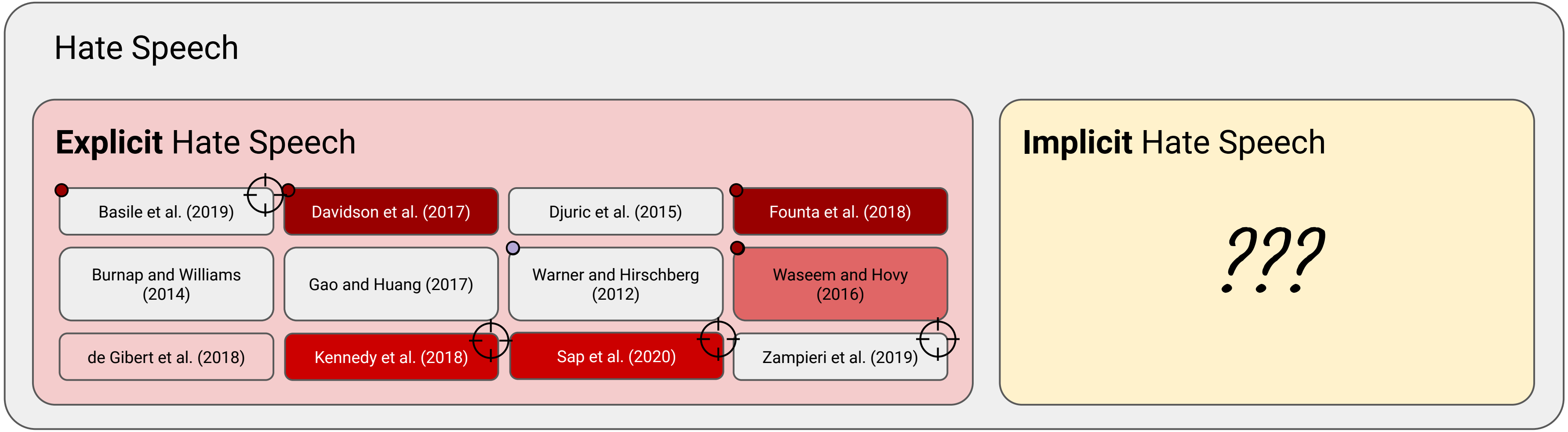
Hate Speech Datasets

Hate Speech

Explicit Hate Speech

- Basile et al. (2019)
- Davidson et al. (2017)
- Djuric et al. (2015)
- Founta et al. (2018)
- Burnap and Williams (2014)
- Gao and Huang (2017)
- Warner and Hirschberg (2012)
- Waseem and Hovy (2016)
- de Gibert et al. (2018)
- Kennedy et al. (2018)
- Sap et al. (2020)
- Zampieri et al. (2019)

Hate Speech Datasets



Hate Speech Datasets

Hate Speech

Explicit Hate Speech

- Basile et al. (2019)
- Davidson et al. (2017)
- Djuric et al. (2015)
- Founta et al. (2018)
- Burnap and Williams (2014)
- Gao and Huang (2017)
- Warner and Hirschberg (2012)
- Waseem and Hovy (2016)
- de Gibert et al. (2018)
- Kennedy et al. (2018)
- Sap et al. (2020)
- Zampieri et al. (2019)

Implicit Hate Speech

- Kennedy et al. (2018)

Hate Speech Datasets

Hate Speech

Explicit Hate Speech

(intentional)

stereotyping, racism, sexism, targeted threats

Implicit Hate Speech

Hate Speech Datasets

Hate Speech

Explicit Hate Speech

(intentional)

stereotyping, racism, sexism, targeted threats

Implicit Hate Speech

(intentionally harmful)

circumlocution, coded language, colloquialisms, connotations, dog whistles, entity framing, euphemisms, hidden threats, idioms, inferiority assumptions, irony, metaphors, presuppositions, symbolic language, ...

Hate Speech Datasets

Hate Speech

Explicit Hate Speech

(intentional)

stereotyping, racism, sexism, targeted threats

Implicit Hate Speech

(intentionally harmful)

circumlocution, coded language, colloquialisms, connotations, dog whistles, entity framing, euphemisms, hidden threats, idioms, inferiority assumptions, irony, metaphors, presuppositions, symbolic language, ...

Bias (e.g. *Social Bias Frames*, *PowerTransformer*)

Microaggressions (Breitfeller et al. 2019)

Hate Speech Datasets

Hate Speech

Explicit Hate Speech

(intentional)

stereotyping, racism, sexism, targeted threats

Implicit Hate Speech

(intentionally harmful)

circumlocution, coded la… …oquialisms, connotations, dog whistles, entity f… …phemisms, hidden threats, idioms, inferiority assumpti… …metaphors, presuppositions, symbolic language, …

[MISSING FROM THE LITERATURE]

Bias (e.g. *Social Bias Frames*, *PowerTransformer*)

Microaggressions (Breitfeller et al. 2019)

Hate Speech Datasets

Hate Speech

Explicit Hate Speech

(intentional)

stereotyping,

racism, sexism,

targeted threats

Impl

(intentionally ha

circumlocution, coded la ... tations, dog

whistles, entity f ... phemisms, ... inferiority

assumpti ... metaphors, presuppositions, ... age, ...

[MISSING ... LITERATURE]

Can we {understand, detect, automatically explain}
this complex subset of hate speech?

Bias (e.g. *Social Bias Frames*, *PowerTransformer*)

Microaggressions (Breitfeller et al. 2019)

Our Contributions

1. Implicit Hate <u>Taxonomy</u>
    a. Incitement of Violence
    b. Inferiority Language
    c. Irony
    d. Stereotypes and Misinformation
    e. Threats and Intimidation
    f. White Grievance

Our Contributions

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset

Our Contributions

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset
   a. hate target*

Our Contributions

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset
    a. hate target*
    b. implied meaning*

Our Contributions

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset
   a. hate target*
   b. implied meaning*
   c. fine-grained implicit hate category

Our Contributions

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset
   a. hate target*
   b. implied meaning*
   c. fine-grained implicit hate category
3. Baseline Classifiers for Detecting Implicit Hate

   transformers

Our Contributions

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset
   a. hate target*
   b. implied meaning*
   c. fine-grained implicit hate category
3. Baseline Classifiers for Detecting Implicit Hate
4. Generative LMs for Explaining Implicit Hate

   GPT / GPT-2

1. Implicit Hate Taxonomy ⚠️

1. Implicit Hate Taxonomy ⚠️

Incitement of Violence

Inferiority Language

Irony

Stereotypes and Misinformation

Threats and Intimidation

White Grievance

1. Implicit Hate Taxonomy ⚠️

> Incitement of Violence 🥊

“Hitler was Germany –Germans shall rise again!”

Inferiority Language

Irony

Stereotypes and Misinformation

Threats and Intimidation

White Grievance

1. Implicit Hate Taxonomy ⚠️

Incitement of Violence 🥊

> Inferiority Language 👎

“It’s not a coincidence the best places to live are majority white.”

Irony

Stereotypes and Misinformation

Threats and Intimidation

White Grievance

1. Implicit Hate Taxonomy ⚠️

Incitement of Violence

Inferiority Language

> Irony

“Horrors... Disney will be forced into hiring Americans”

Stereotypes and Misinformation

Threats and Intimidation

White Grievance

1. Implicit Hate Taxonomy ⚠️

Incitement of Violence 🥊

Inferiority Language 👎

Irony 🎭

> Stereotypes and Misinformation ⚖️

“Can someone tell the black people in Chicago to stop killing one another before it becomes Detroit?”

Threats and Intimidation

White Grievance

1. Implicit Hate Taxonomy ⚠️

Incitement of Violence

Inferiority Language

Irony

Stereotypes and Misinformation

> Threats and Intimidation

“It won't be long before whitey is walking thru the slums, hoping and preying for a white victory.”

White Grievance

1. Implicit Hate Taxonomy ⚠️

Incitement of Violence 🥊

Inferiority Language 👎

Irony 🎭

Stereotypes and Misinformation ⚖️

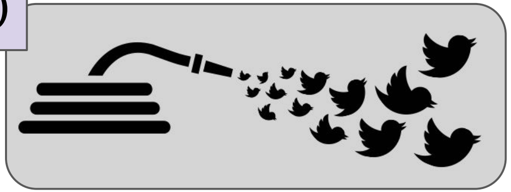
Threats and Intimidation 🔫

> White Grievance 🌎

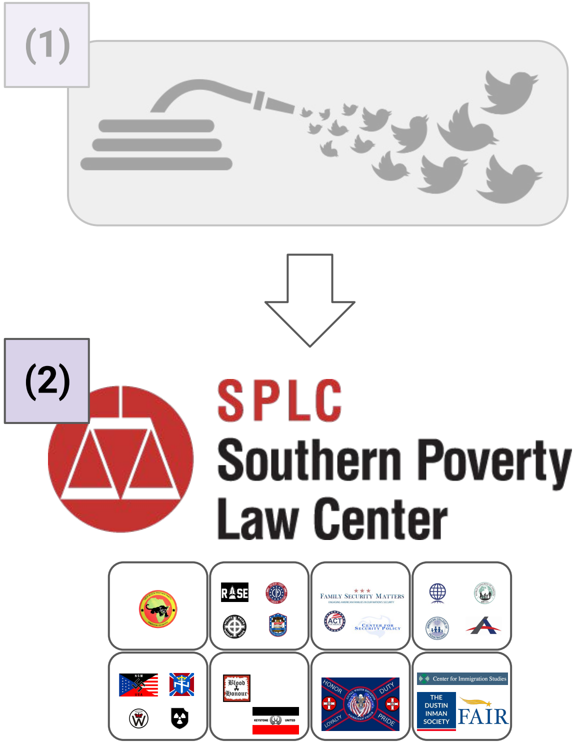
“Black lives matter and white lives don’t? Sounds racist.”

2. Implicit Hate Benchmark Dataset

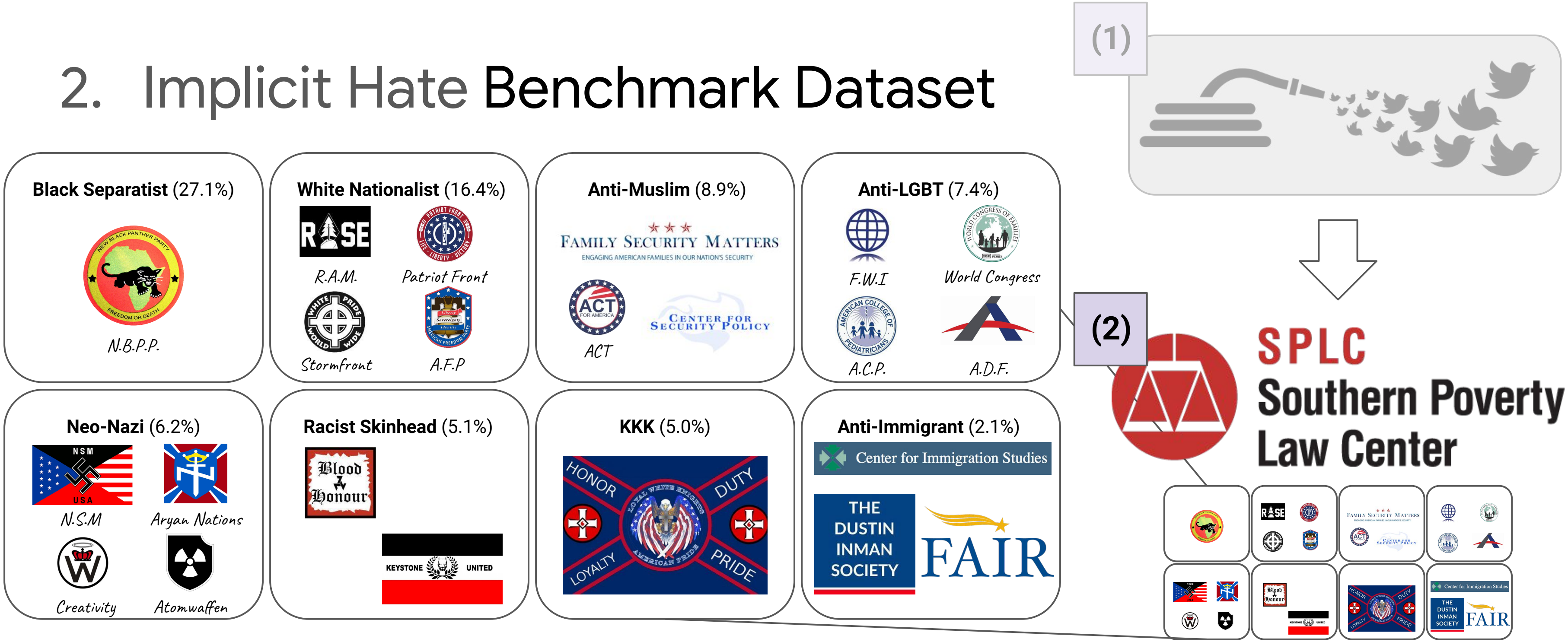
(1)



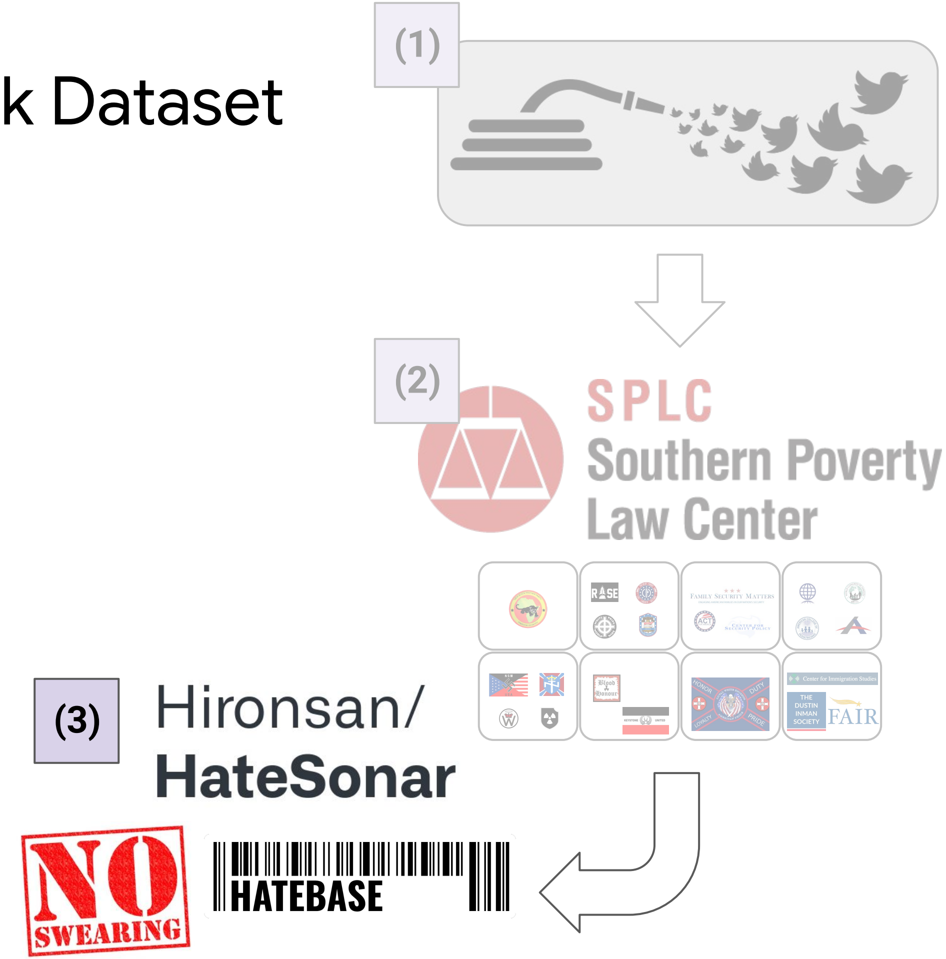
2. Implicit Hate Benchmark Dataset



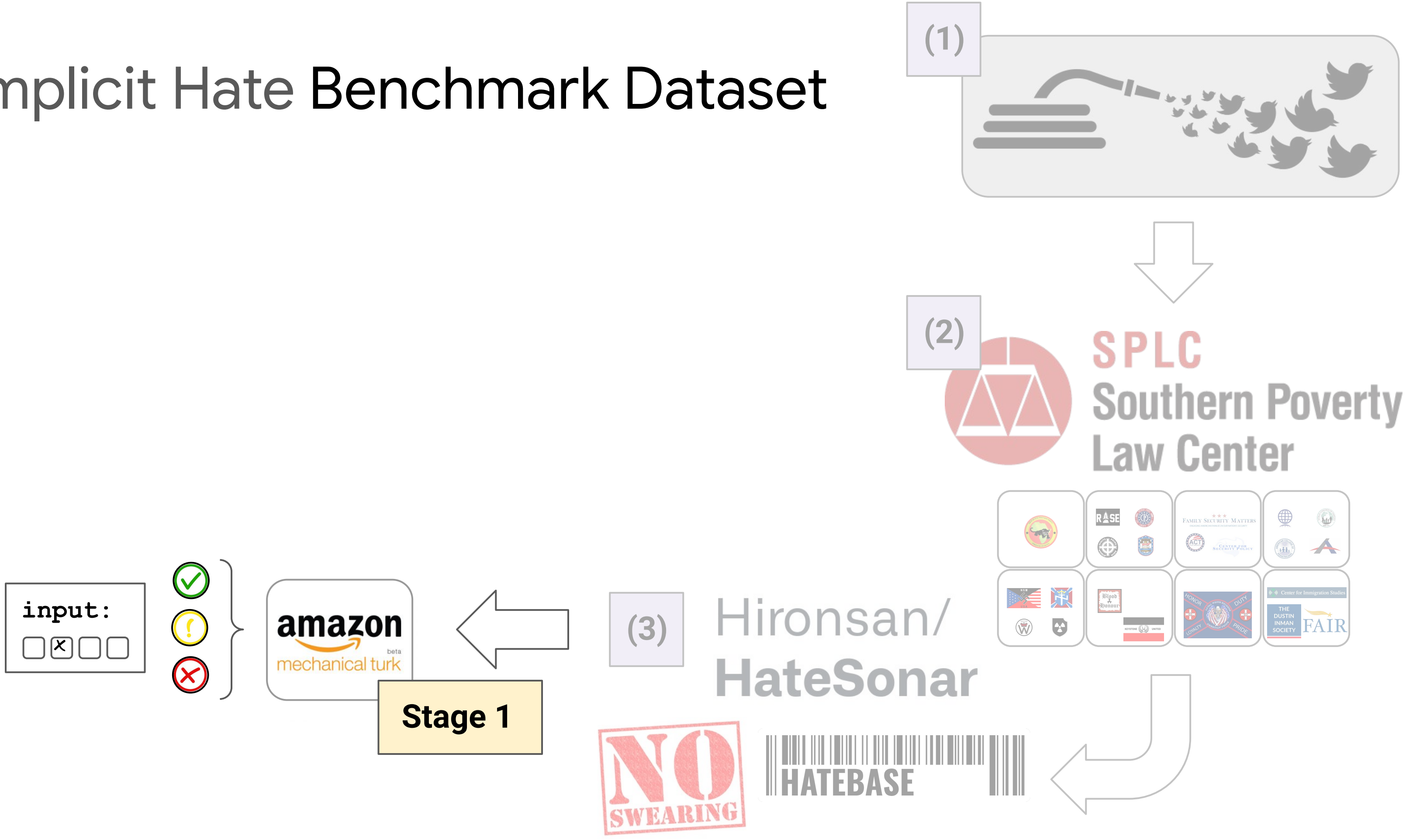
2. Implicit Hate Benchmark Dataset



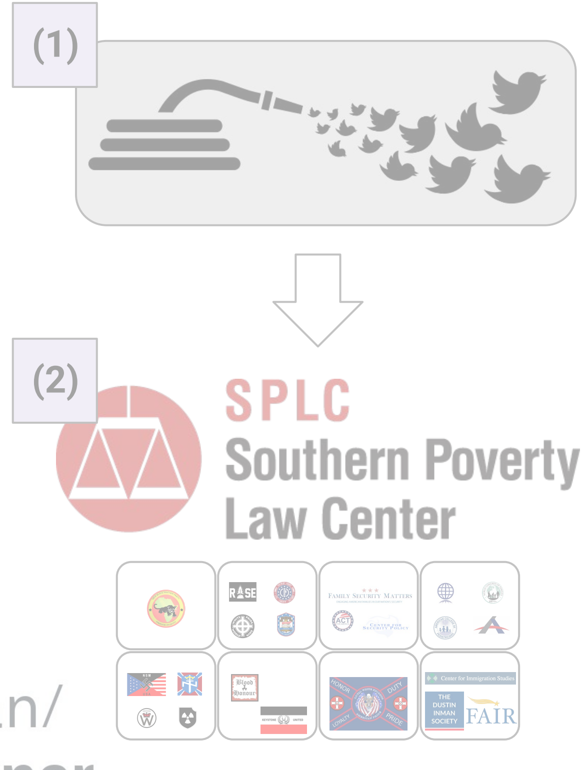
2. Implicit Hate Benchmark Dataset

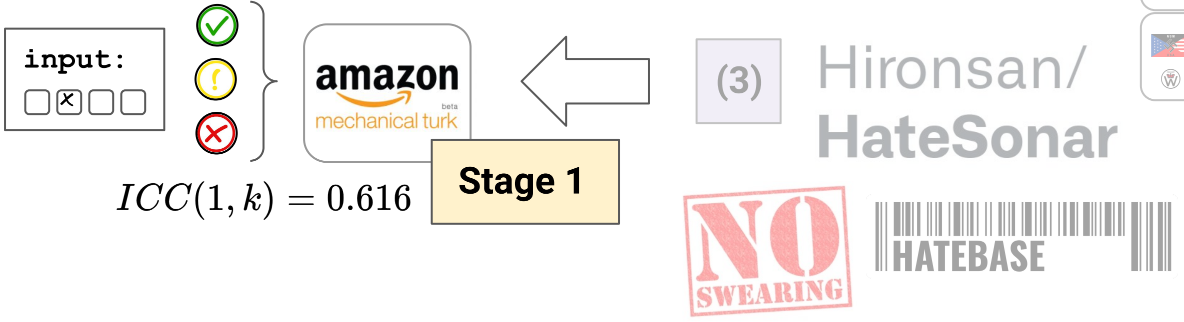


2. Implicit Hate Benchmark Dataset

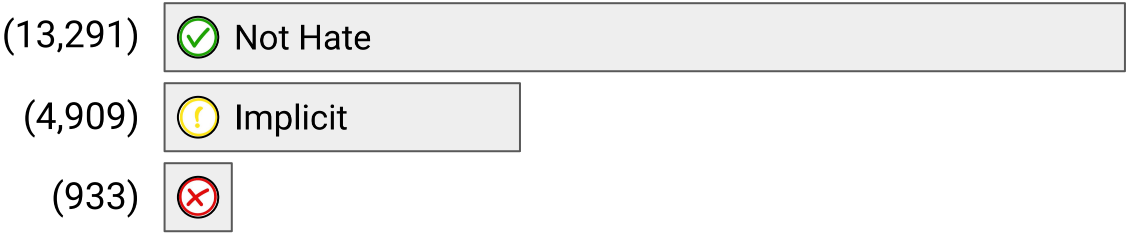


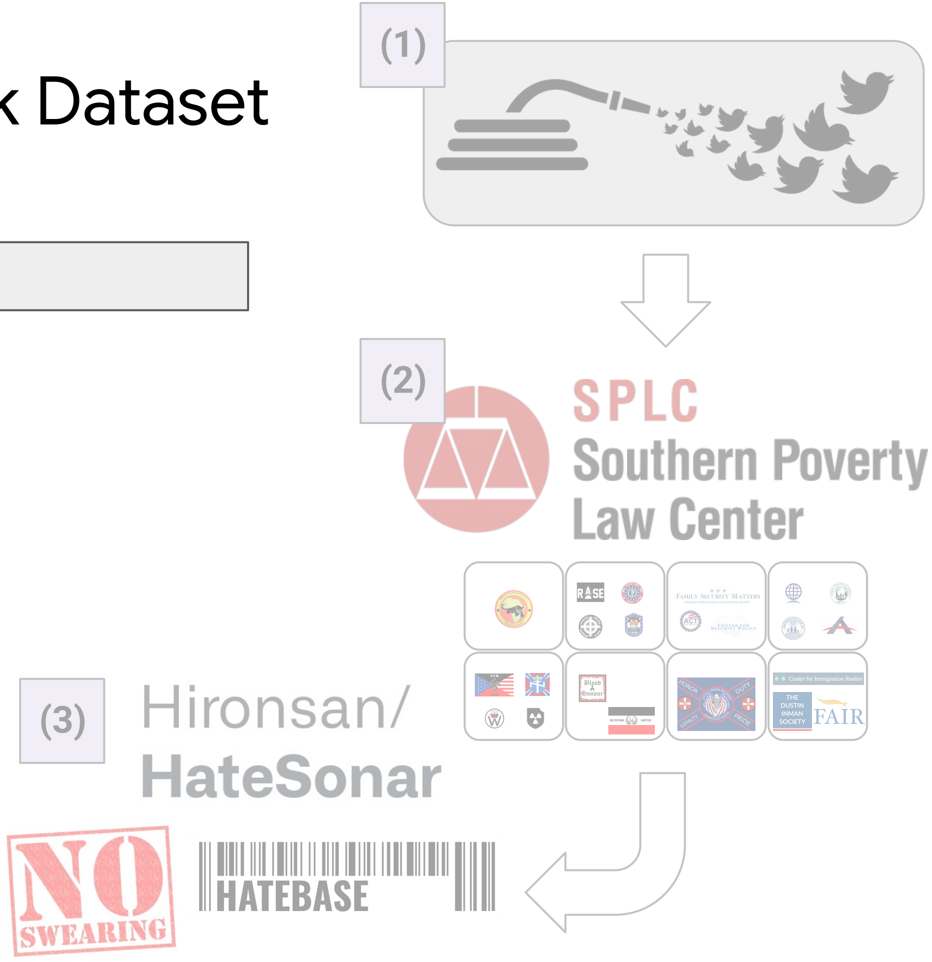
2. Implicit Hate Benchmark Dataset

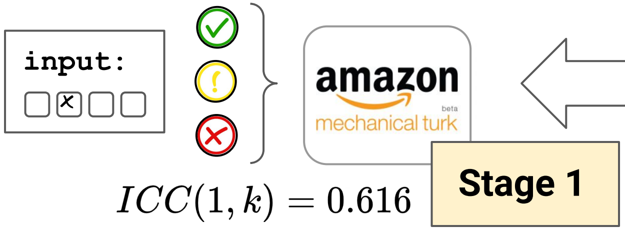




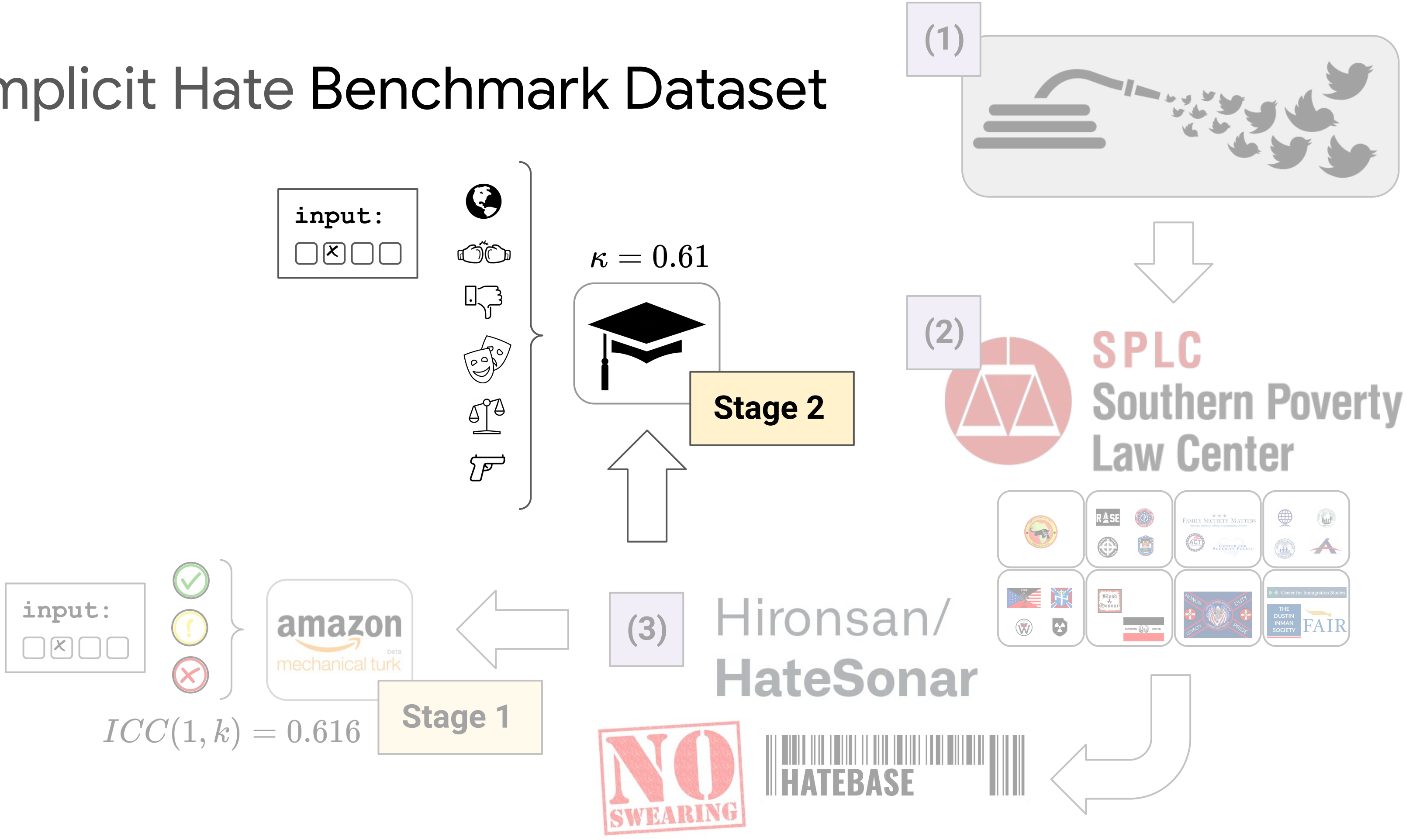
2. Implicit Hate Benchmark Dataset



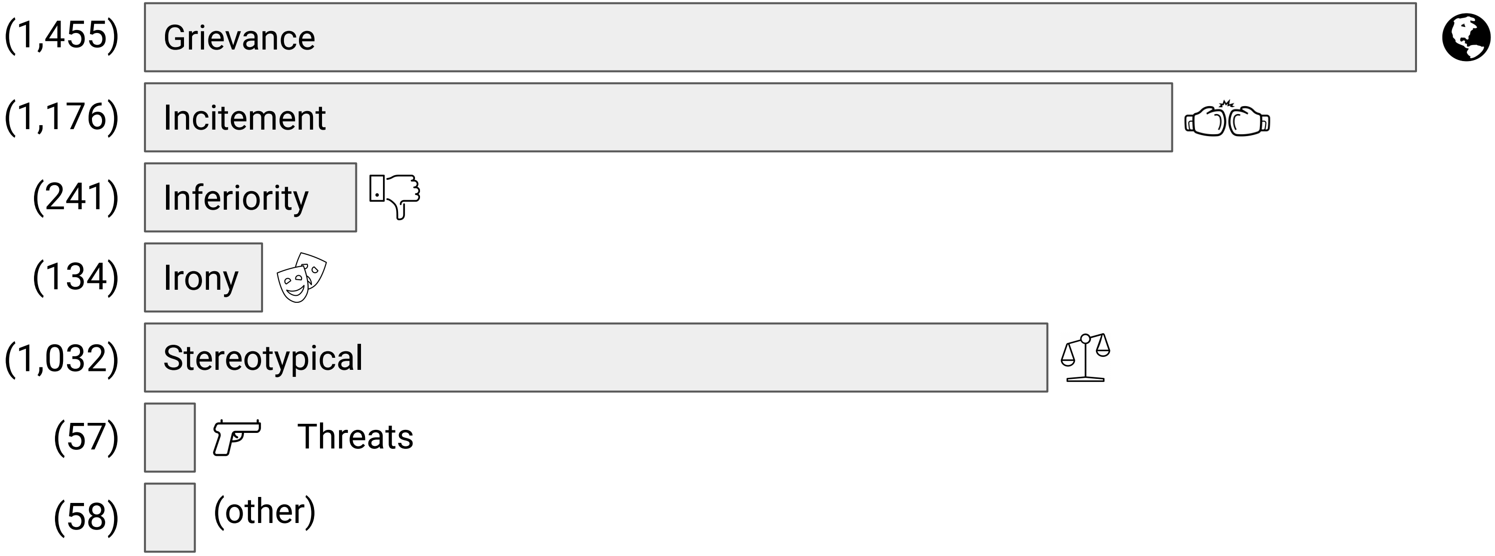




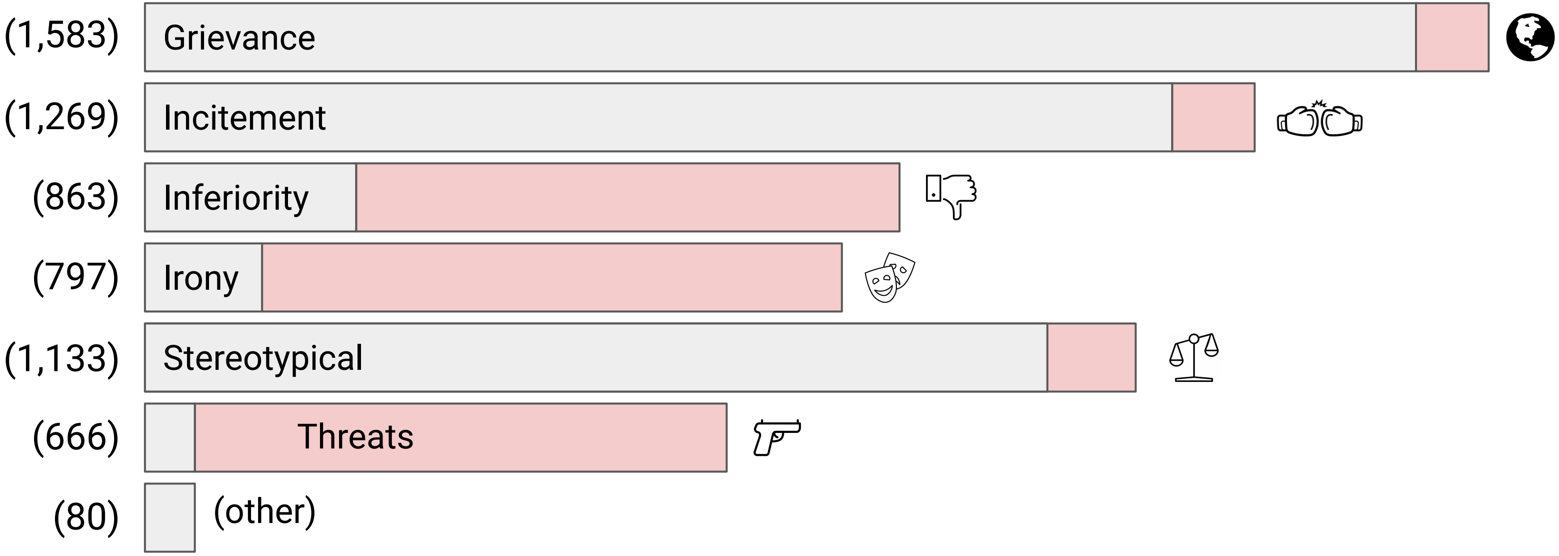
2. Implicit Hate Benchmark Dataset



2. Implicit Hate Benchmark Dataset

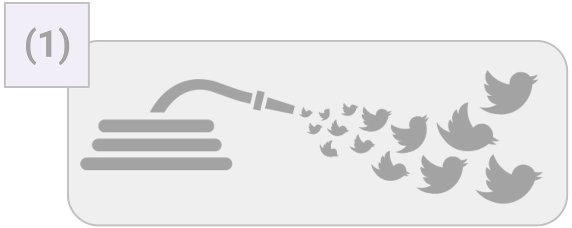


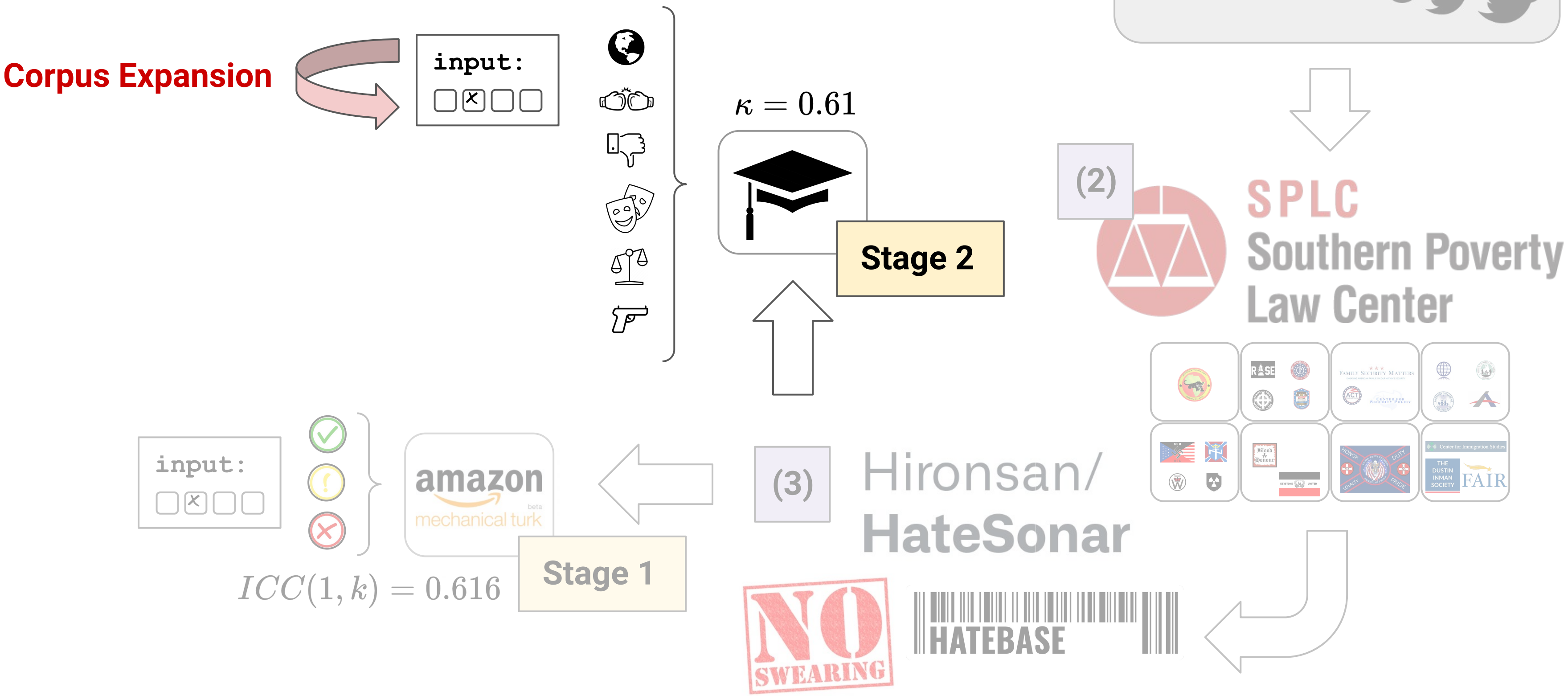
2. Implicit Hate Benchmark Dataset



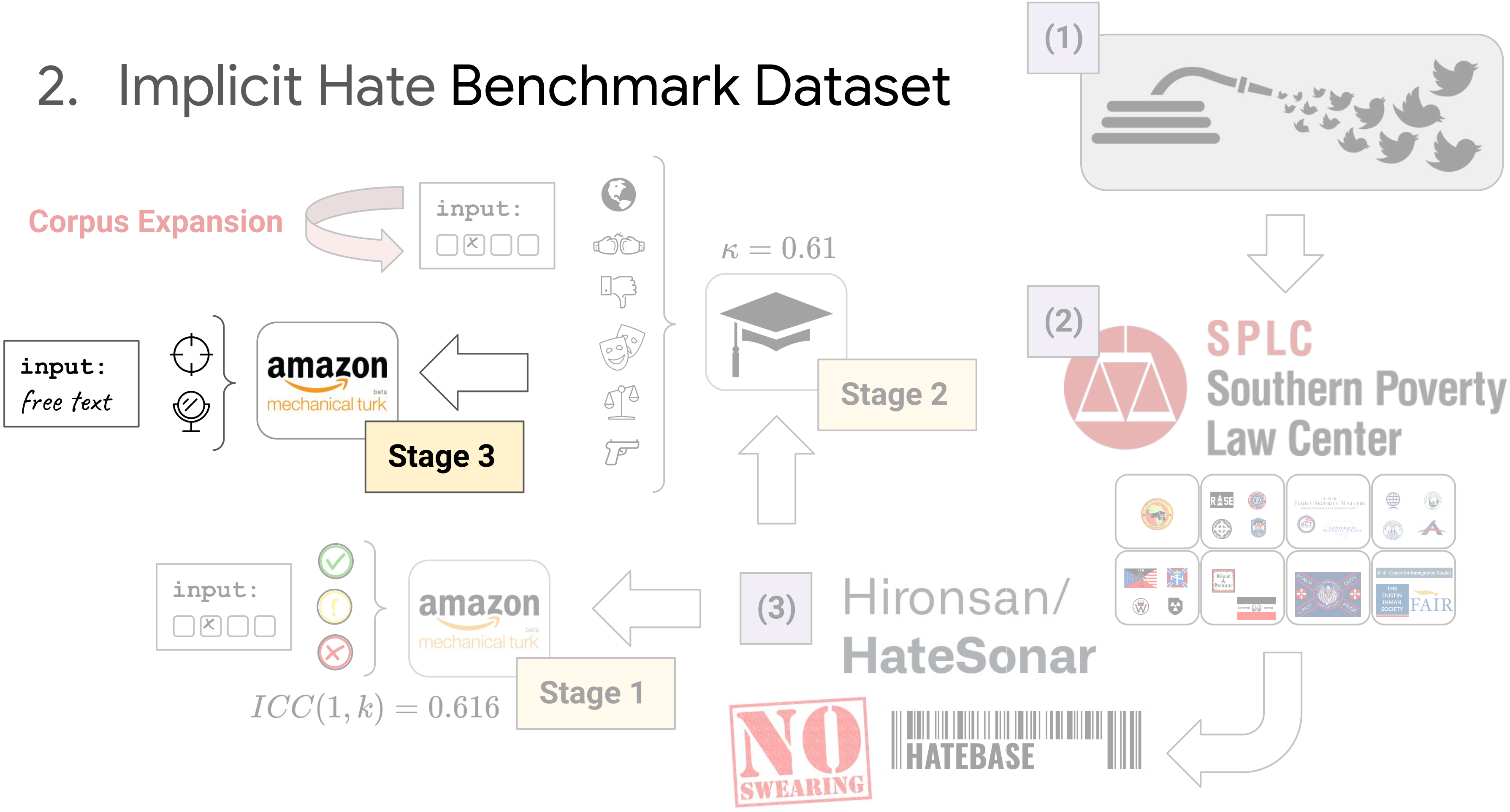
6,346 Total After **Corpus Expansion**

2. Implicit Hate Benchmark Dataset

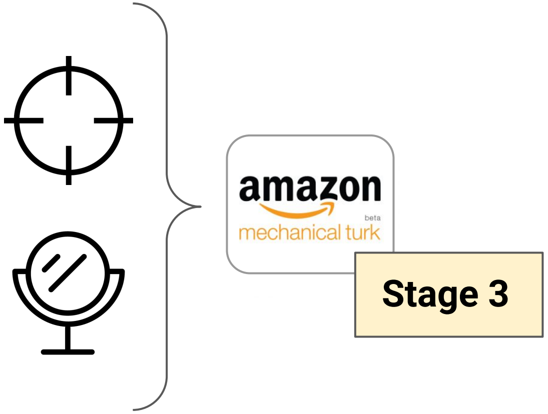




2. Implicit Hate Benchmark Dataset



2. Implicit Hate Benchmark Dataset



2. Implicit Hate Benchmark Dataset ⚠️

Target: {Black folks, Muslim folks, Non-whites}

Implied Statement: `<targets>` {do, are, commit} `<predicate>`

e.g. "*Mexicans* are *incompetent*"

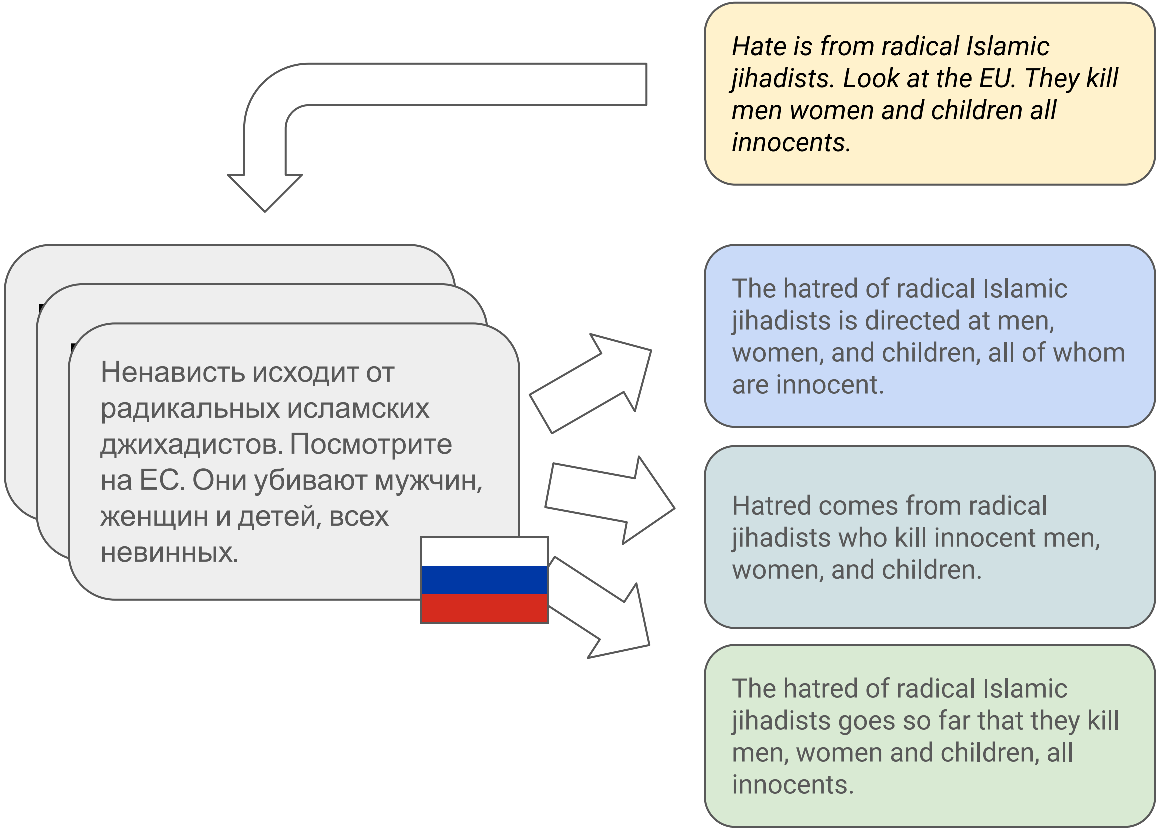
3. Baseline Classifiers for Detecting Implicit Hate ⚠️

SVM (n-grams)
SVM (TF-IDF)
SVM (GloVe)
BERT
BERT + Aug
BERT + Aug + Wikidata
BERT + Aug + ConceptNet

3. Baseline Classifiers for Detecting Implicit Hate ⚠️

Augmentation

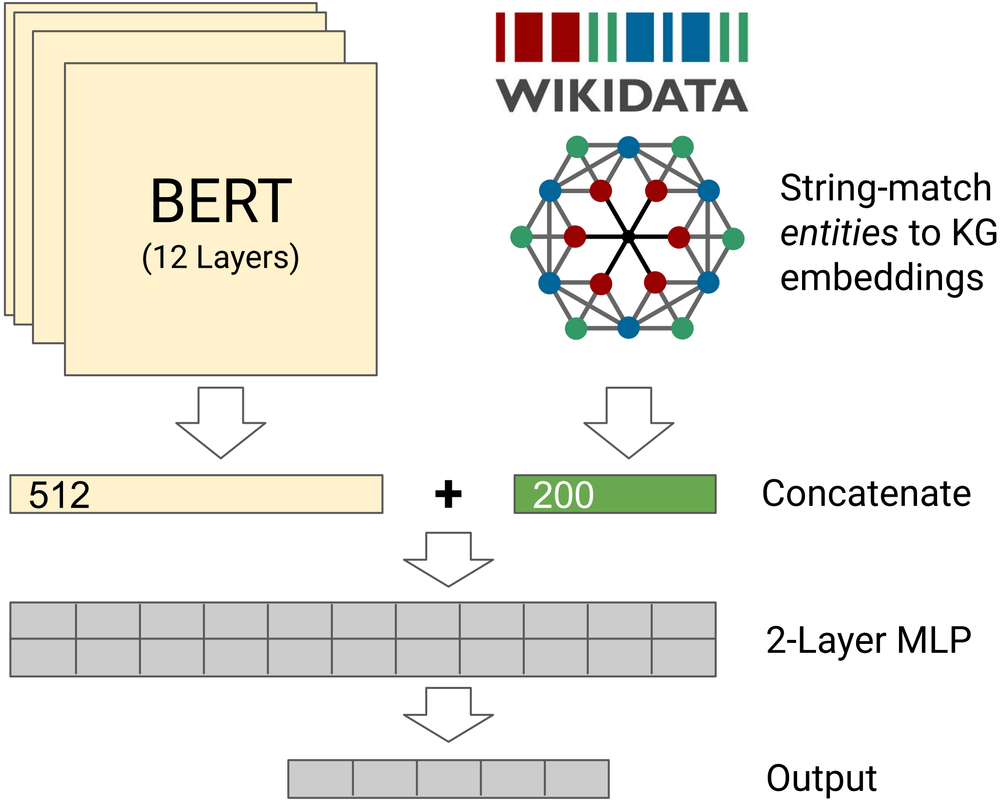
SVM (n-grams)
SVM (TF-IDF)
SVM (GloVe)
BERT
$\boxed{\text{BERT + Aug}}$
BERT + Aug + Wikidata
BERT + Aug + ConceptNet



3. Baseline Classifiers for Detecting Implicit Hate

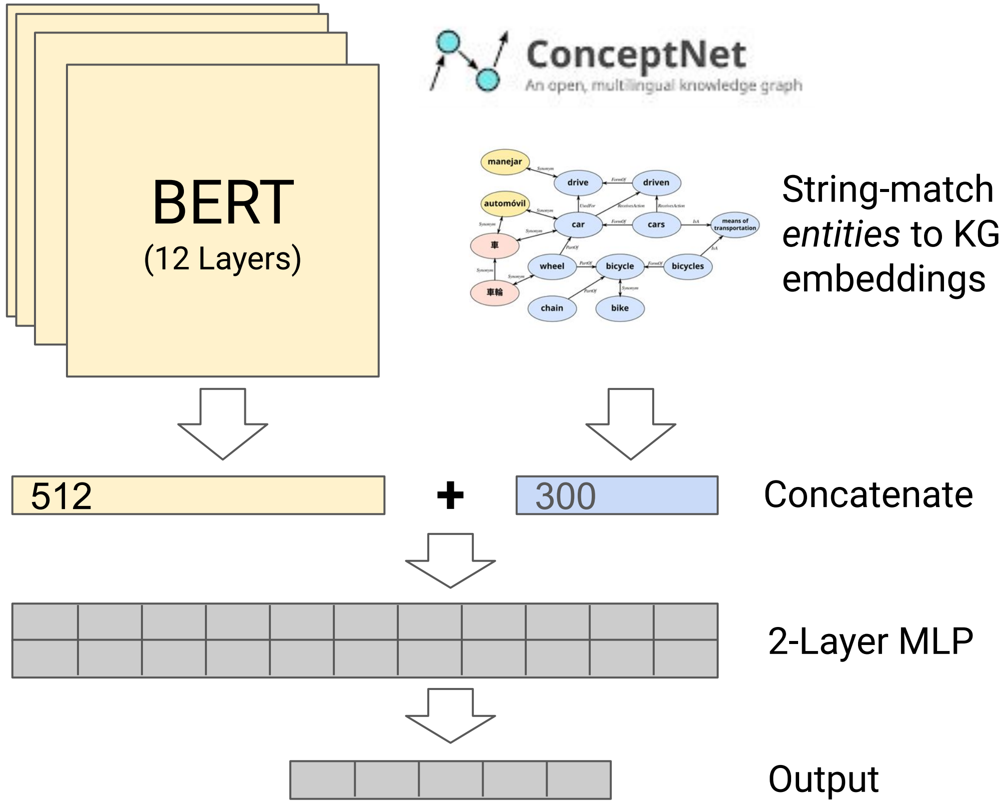
WikiData

SVM (n-grams)
SVM (TF-IDF)
SVM (GloVe)
BERT
BERT + Aug
BERT + Aug + Wikidata
BERT + Aug + ConceptNet



3. Baseline Classifiers for Detecting Implicit Hate

ConceptNet
SVM (n-grams)
SVM (TF-IDF)
SVM (GloVe)
BERT
BERT + Aug
BERT + Aug + Wikidata
BERT + Aug + ConceptNet



3. Baseline Classifiers for Detecting Implicit Hate

Models
SVM (n-grams)
SVM (TF-IDF)
SVM (GloVe)
BERT
BERT + Aug
BERT + Aug + Wikidata
BERT + Aug + ConceptNet

3. Baseline Classifiers for Detecting Implicit Hate

	Binary Classification				**Implicit Hate Categories**			
Models	P	R	F	Acc	P	R	F	Acc
Hate Sonar	39.9	48.6	43.8	54.6	-	-	-	-
Perspective API	50.1	61.3	55.2	63.7	-	-	-	-
SVM (n-grams)	61.4	67.7	64.4	72.7	48.8	49.2	48.4	54.2
SVM (TF-IDF)	59.5	68.8	63.9	71.6	53.0	51.7	51.5	56.5
SVM (GloVe)	56.5	65.3	60.6	69.0	46.8	48.9	46.3	51.3
BERT	**72.1**	66.0	68.9	**78.3**	**59.1**	57.9	58.0	62.9
BERT + Aug	67.8	**73.2**	**70.4**	77.5	58.6	**59.1**	**58.6**	**63.8**
BERT + Aug + Wikidata	67.6	72.3	69.9	77.3	53.9	55.3	54.4	62.8
BERT + Aug + ConceptNet	68.6	70.0	69.3	77.4	54.0	55.4	54.3	62.5

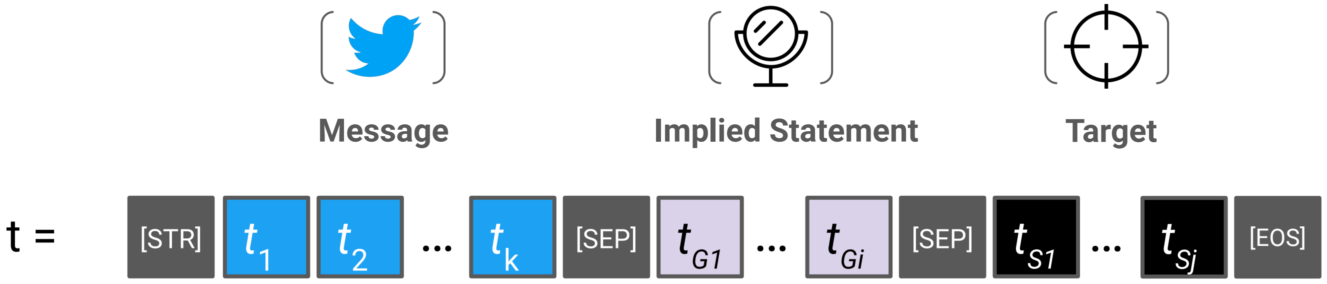
3. Baseline Classifiers for Detecting Implicit Hate

Error Analysis

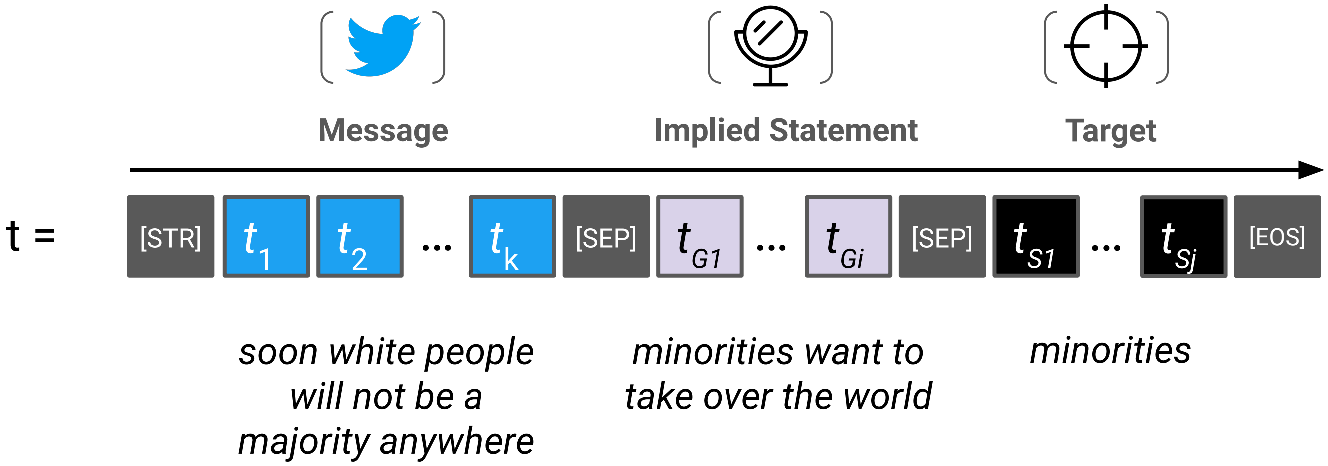
- Coded Hate Symbols
- Entity framing
- Metaphorical language
- Irony
- Discourse relations
- Commonsense
- Colloquial speech
- Identity-term bias

4. Generative LMs for Explaining Implicit Hate

4. Generative LMs for Explaining Implicit Hate

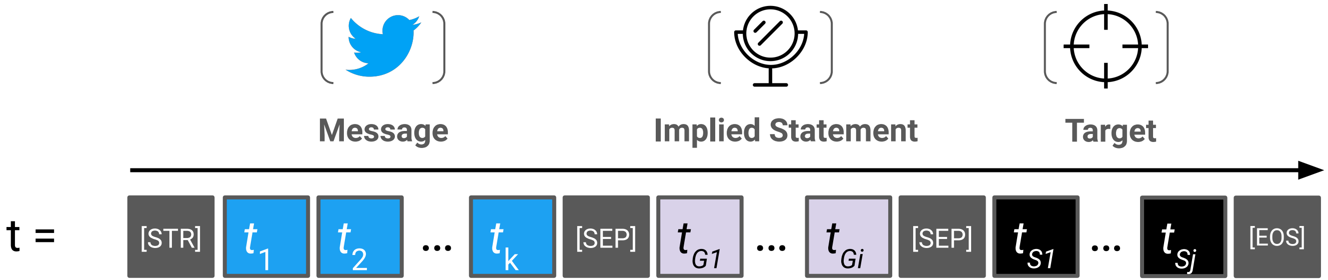


4. Generative LMs for Explaining Implicit Hate



Cross-Entropy Loss: $-\sum_l \log P(\tilde{t}_l | t_{<l})$

4. Generative LMs for **Explaining Implicit Hate**



soon white people will not be a majority anywhere

minorities want to take over the world

minorities

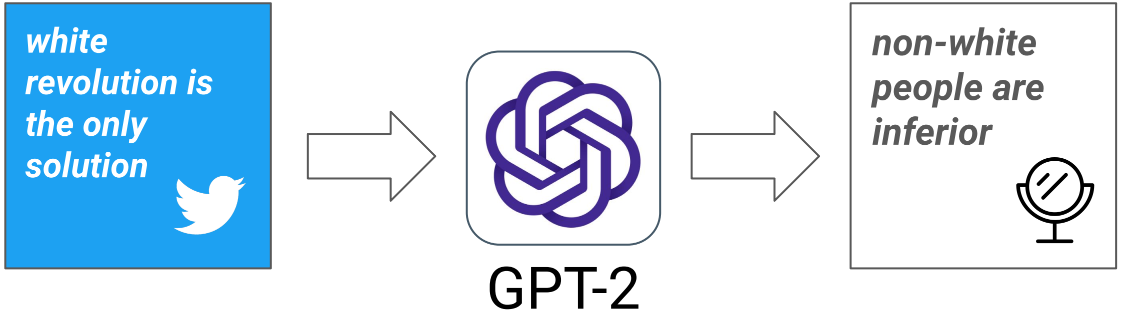
Cross-Entropy Loss: $-\sum_l \log P(\tilde{t}_l | t_{<l})$

Models: GPT GPT-2

4. Generative LMs for Explaining Implicit Hate

	Target Group				Implied Statement			
Models	BLEU	BLEU*	Rouge-L	Rouge-L*	BLEU	BLEU*	Rouge-L	Rouge-L*
GPT-gdy	43.7	65.2	42.9	63.3	41.1	58.2	31	45.3
GPT-top-p	57.7	76.8	55.8	74.6	55.2	69.4	40	53.9
GPT-beam	59.3	81	57.3	78.6	57.8	73.8	46.5	63.4
GPT-2-gdy	45.3	67.6	44.6	66	42.3	59.3	32.7	47.4
GPT-2-top-p	58.0	76.9	56.2	74.8	55.1	69.3	39.6	53.1
GPT-2-beam	**61.3**	**83.9**	**59.6**	**81.8**	**58.9**	**75.3**	**48.3**	**65.9**

4. Generative LMs for Explaining Implicit Hate

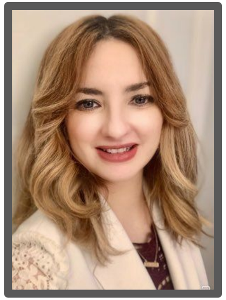


ground truth:
white people are superior to other races

Conclusion

1. Implicit Hate Taxonomy
2. Implicit Hate Benchmark Dataset
3. Baseline Classifiers for Detecting Implicit Hate
4. Generative LMs for Explaining Implicit Hate

GT-SALT/
implicit-hate



@mai_elsherief

@cjziems

@DMuchlinski

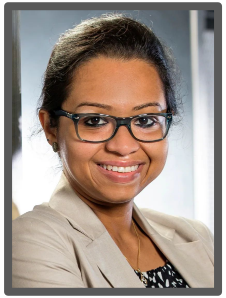
Vaishnavi Anupindi

Jordyn Seybolt

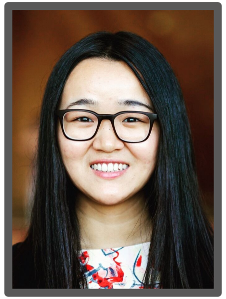
Thanks!

Please reach out

GT-SALT/
implicit-hate



@munmun10



@Diyi_Yang

Post	**Type**	**Target Group**	**Implied Statement**
(a) white revolution is the only solution	incitement	non-white people *white people*	non-white people are inferior *white people are superior to other races*
(b) yes you are fine in a white majority country. how is mexico doing these days? people come here because you couldn't build it.	inferiority	mexican people *mexicans*	mexican people do not build things *mexicans are incompetent*
(c) black people only seem to thrive in white countries. how strange	irony	black people *black people*	black people are inferior *black people can't survive without white people*
(d) soon white people will not be a majority anywhere. you believe	grievance	whites *minorities*	whites have been driven out of the country *minorities want to take over the world*
(e) muslims are always wanting to kill someone! its their duty to kill the infidel.	stereotypical	muslims *muslims*	muslims are killers *muslims are violent*
(f) oh how i wish a man with a gun was there! wake up white people! they hate us	threatening	white people *non-whites*	white people are hated *non-whites hate whites*

Table 5: Example posts from our dataset along with their implicit category labels, the GPT-2 generated target and implied statements (first row of each block), and the ground truth target and implied statements (final row of each block, in *italics*). Generated implied statements are semantically similar to the ground truth statements.

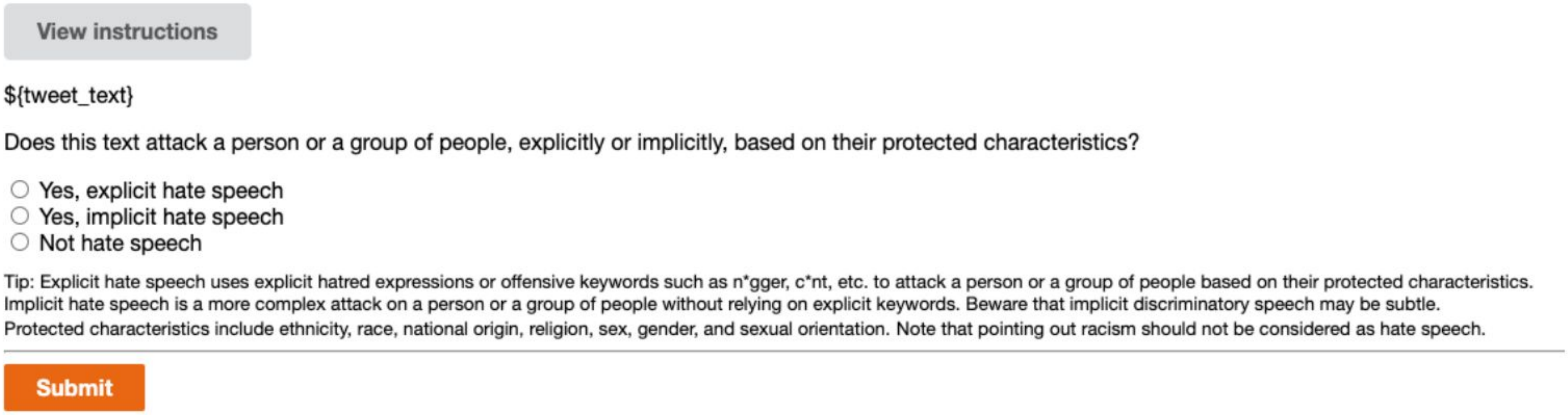
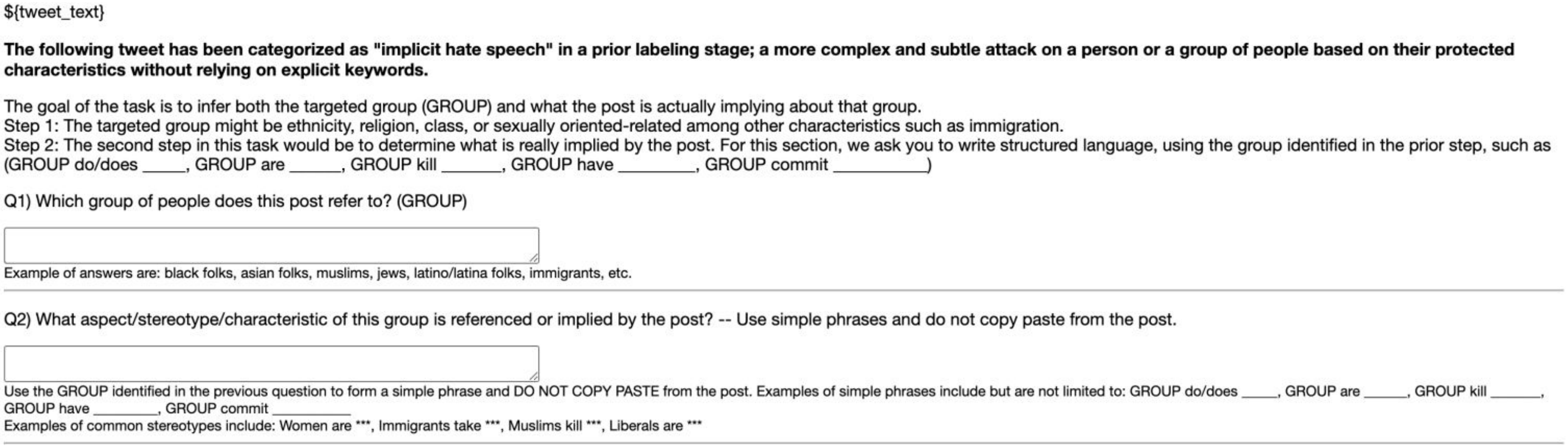


Figure 2: Amazon Mechanical Turk interface used to collect ternary annotations (explicit hate, implicit hate, and not hate) for our first stage.



${tweet_text}

The following tweet has been categorized as "implicit hate speech" in a prior labeling stage; a more complex and subtle attack on a person or a group of people based on their protected characteristics without relying on explicit keywords.

The goal of the task is to infer both the targeted group (GROUP) and what the post is actually implying about that group.
Step 1: The targeted group might be ethnicity, religion, class, or sexually oriented-related among other characteristics such as immigration.
Step 2: The second step in this task would be to determine what is really implied by the post. For this section, we ask you to write structured language, using the group identified in the prior step, such as (GROUP do/does _____, GROUP are ______, GROUP kill _______, GROUP have ________, GROUP commit __________)

Q1) Which group of people does this post refer to? (GROUP)

Example of answers are: black folks, asian folks, muslims, jews, latino/latina folks, immigrants, etc.

Q2) What aspect/stereotype/characteristic of this group is referenced or implied by the post? -- Use simple phrases and do not copy paste from the post.

Use the GROUP identified in the previous question to form a simple phrase and DO NOT COPY PASTE from the post. Examples of simple phrases include but are not limited to: GROUP do/does _____, GROUP are ______, GROUP kill _______, GROUP have ________, GROUP commit __________
Examples of common stereotypes include: Women are ***, Immigrants take ***, Muslims kill ***, Liberals are ***

Figure 3: Amazon Mechanical Turk interface used to collect the hate target and the implied statement per implicit hate speech post.

	Macro				Grievance			Incitement			Inferiority		
	P	R	F	Acc	P	R	F	P	R	F	P	R	F
SVM (n-grams)	48.8	49.2	48.4	54.2	65.6	53.6	59.0	53.7	55.8	54.7	49.7	46.4	48.0
SVM (TF-IDF)	53.0	51.7	51.5	56.5	66.9	56.7	61.4	60.4	56.2	58.2	46.0	45.3	45.6
SVM (GloVe)	46.8	48.9	46.3	51.3	63.7	48.6	55.1	55.2	46.7	50.6	45.8	39.7	42.5
BERT	**59.1**	57.9	58.0	62.9	65.4	63.9	64.6	62.4	**56.6**	59.4	**65.4**	57.9	61.4
BERT + Aug	58.6	**59.1**	**58.6**	**63.8**	67.6	**65.7**	**66.6**	**66.8**	56.5	**61.2**	61.0	59.0	59.9
BERT + Aug + Wikidata	53.9	55.3	54.4	62.8	**68.8**	63.0	65.8	62.7	55.9	59.1	60.3	**60.8**	**60.4**
BERT + Aug + ConceptNet	54.0	55.4	54.3	62.5	67.6	64.9	66.2	63.8	52.7	57.7	62.1	57.7	59.7

	Irony			Stereotypical			Threatening		
	P	R	F	P	R	F	P	R	F
SVM (n-grams)	41.4	51.8	46.0	60.7	52.7	56.4	52.0	72.2	60.5
SVM (TF-IDF)	43.9	55.4	48.9	60.9	58.8	59.8	55.3	72.2	62.7
SVM (GloVe)	48.7	55.4	51.8	59.3	53.9	56.5	50.2	74.3	59.9
BERT	**62.3**	**63.8**	**63.0**	58.5	69.3	63.4	**67.2**	71.5	69.3
BERT + Aug	62.0	62.3	62.1	**62.0**	**70.1**	**65.8**	65.0	**75.6**	**69.8**
BERT + Aug + Wikidata	60.0	63.1	61.4	60.7	69.3	64.7	64.2	73.8	68.6
BERT + Aug + Conceptnet	61.5	63.3	62.3	59.1	70.0	64.0	62.4	74.7	67.9

Table 6: Fine-grained implicit hate classification performance, averaged across five random seeds. Macro scores are further broken down into category-level scores for each of the six main implicit categories, and we omit scores for *other*. Again, the BERT-based models beat the linear SVMs on F_1 performance across all categories. Generally, augmentation improves recall, especially for two of the minority classes, *inferiority* and *threatening*, as expected. Knowledge graph integration (Wikidata, Conceptnet) does not appear to improve the performance.

	White Nationalist	Neo-Nazi	A-Immgr	A-MUS	A-LGBTQ	KKK
Nouns (N)	identity evropa activists alt-right whites	adolf bjp india modi invaders	immigration sanctuary aliens border cities	islam jihad islamic muslim(s) sharia	potus democrats trump abortion dumbocrats	ku klux hood niggas brother
Adjectives (A)	white hispanic anti-white third racial	more non-white german national-socialist white	illegal immigrant dangerous ice criminal	muslim political islamic migrant moderate	black crooked confederate fake racist	alive edgy white outed anonymous
Hashtags (#)	#projectsiege #antifa #berkrally #altright #endimmigration	#swrm #workingclass #hitler #freedom #wpww	#noamnesty #immigration #afire #fairblog #stopsanctuarycities	#billwarnerphd #stopislam #makedclisten #banssharia #cspi	#defundpp #pjnet #unbornlivesmatter #religiousfreedom #prolife	#opkkk #hoodsoff #mantears #kkk #anonymous

Table 7: Top five salient nouns, adjectives, and hashtags identified by measuring the log odds ratio informative Dirichlet prior (Monroe et al., 2008) for the following ideologies: White Nationalist, Neo-Nazi, Anti-Immigrant (A-Immgr), Anti-Muslim (A-MUS), Anti-LGBTQ (A-LGBTQ), and Ku Klux Klan (KKK).

2. Implicit Hate Benchmark Dataset