Hate Speech Detection *WiSe 23-24*
Implicit Hate Speech

Image taken from https://deepsense.ai/artificial-intelligence-hate-speech/

(Dr.) Özge Alaçam

Computational Linguistics

oezge.alacam@lmu.de

Implicit Hate Speech Detection

- ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). *Latent Hatred: A Benchmark for Understanding Implicit Hate Speech*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

- Kim, Y., Park, S., & Han, Y. S. (2022, October). *Generalizable implicit hate speech detection using contrastive learning*. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6667-6679).

- Jafari, A. R., Li, G., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2023). *Fine-grained emotions influence on implicit hate speech detection*. *IEEE Access*.

- Holt, F., Nguyen, C., & Shah, P. *A Study In Hate: Dissecting Transformer-Based Models' Rationale for Implicit Hate Classification*.

Implicit Hate Speech Dataset

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, †Georgia Institute of Technology
melsherief@ucsd.edu
{cziems, dmuchlinski3, vanupindi3}@gatech.edu
{jseybolt3, munmund, dyang888}@gatech.edu

Contributions of the paper

- a large and representative sample of implicit hate speech with fine-grained implicit hate labels
- natural language descriptions of the implied aspects for each hateful message.
- competitive baseline classifiers to detect implicit hate speech and generate its implied statements.
- >>> attempts to establish a theoretical framework for implicit hate speech

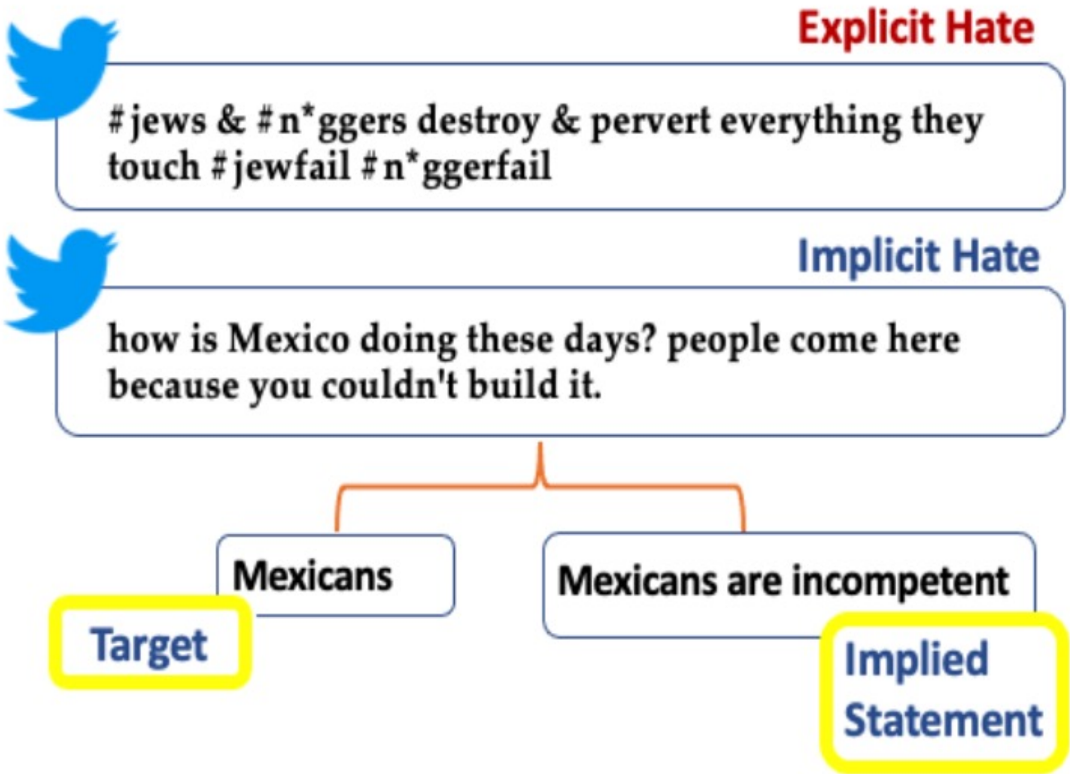


Figure 1: Sample posts from our dataset outlining the differences between explicit and implicit hate speech. Explicit hate is **direct** and leverages specific keywords while implicit hate is more **abstract**. Explicit text has been modified to include a star (*).

- the linguistic nuance and diversity of the implicit hate class, which includes indirect sarcasm and humor

Implicit Hate Taxonomy (not mutually exclusive)

- **White Grievance** includes frustration over a minority group's perceived privilege and casting majority groups as the real victims of racism .
  - Black lives matter and white lives don't? Sounds racist.
- **Incitement to Violence** includes flaunting in- group unity and power or elevating known hate groups and ideologies (Somerville, 2011).
  - Phrases like 'white brotherhood'
- **Inferiority Language** implies one group or individual is inferior to another and it can include dehumanization (denial of a person's humanity), and toxification
  - It's not a coincidence the best places to live are majority white.

Implicit Hate Taxonomy (not mutually exclusive)

- **Irony** refers to the use of sarcasm, humor, and satire to attack or demean a protected class or individual.
  - Horrors… Disney will be forced into hiring Americans

- **Stereotypes and Misinformation** associate a protected class with negative attributes such as crime or terrorism
  - Can someone tell the black people in Chicago to stop killing one another before it becomes Detroit?

- **Threatening and Intimidation** convey a speaker commitment to a target's pain, injury, damage, loss, or violation of rights. (implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation)
  - All immigration of non-whites should be ended.

Annotation Scheme

- Almost 5M tweets from hate group accounts
- First, using MTURK, they collected high-level labels
  - explicit hate, implicit hate, or not hate.
  - Using the majority vote, 933 explicit hate, 4,909 implicit hate, and 13,291 not hateful tweets.
- Then, a second pass through the implicit hate tweets with expert annotation (3 experts highly-trained on the task)
  - over the fine-grained implicit hate taxonomy (6 categories).

Models

- standard unigrams, TF-IDF, and Glove embeddings with SVM
- BERT models
  - Bert-base
  - Bert-base + augmented data (bootstrapping to counter imbalance)

  Incorporating knowledge-based features (adding entity embeddings)
  - Bert-base + aug + wikidata (Wikidata Knowledge Graph)
  - Bert-base +aug+ ConceptNet (Concept Embeddings)

	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

Complexity of Implicit Categories

- Based on the best model scores,

- Most difficult implicit category is Incitement (36.3% of testing examples were classified as not hate),
- followed by White Grievance (29.6%),
- Stereotypical (23.3%),
- Inferiority (12.3%),
- Irony (9.3%),
- Threatening (5.5%).

Next Step: Explaining Implicit Hate Speech

- Generating explanations of
- (1) who is being targeted
- (2) What is the implied message for each implicitly hateful tweet?

- Conditional Generation Task:

  given a post, generate a hateful post's intended target and hidden implied meanings.

Using several decoding strategies such as greedy search (gdy), beam search, and top-p (nucleus) sampling to generate the explanations...

[A really nice guide about text generation decoding strategies can be found here]()

Models	Target Group BLEU	BLEU*	Rouge-L	Rouge-L*	Implied Statement 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**

Table 4: Evaluation of the generation models for Target Group and Implied Statement. (*) denotes the maximum versus the average score (without asterisk). gdy: greedy decoding, beam: beam search with 3 hypotheses, and top-p: nucleus sampling with $p = 0.92$

Example for Inferiority - Implicit Hate Category:

- “yes you are fine in a white majority country. how is mexico doing these days? people come here because you couldn’t build it.”

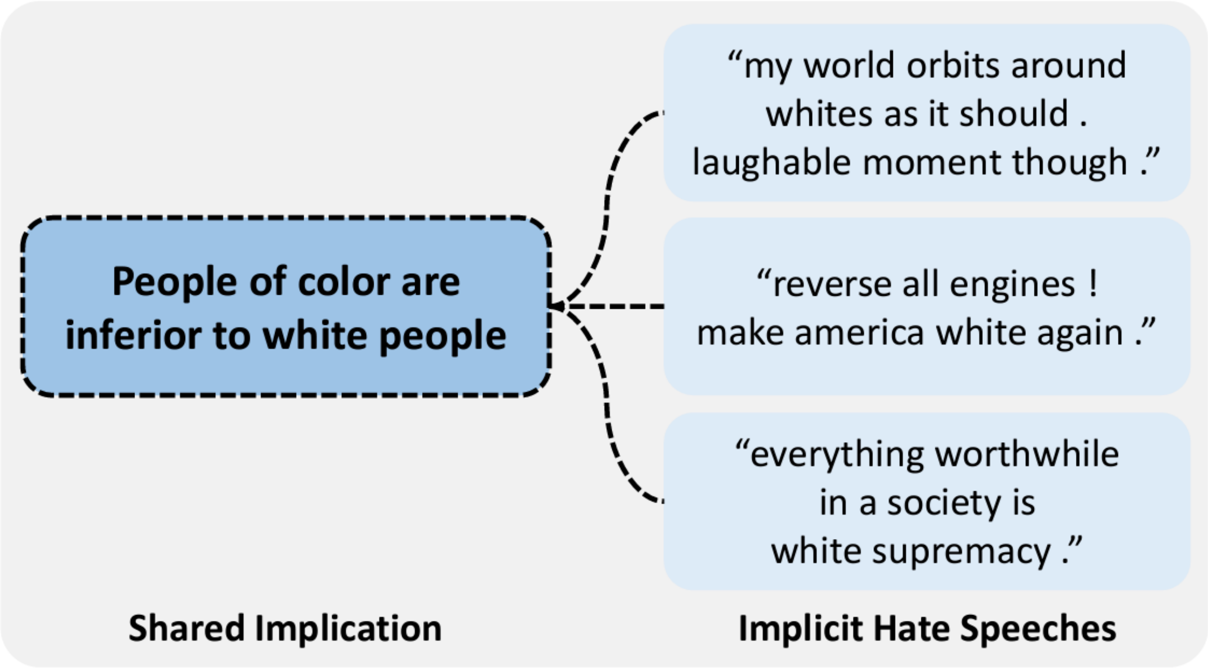
	Target	**Implication**
GPT-2	mexican people	mexican people do not build things
Human Annotator	mexicans	mexicans are incompetent

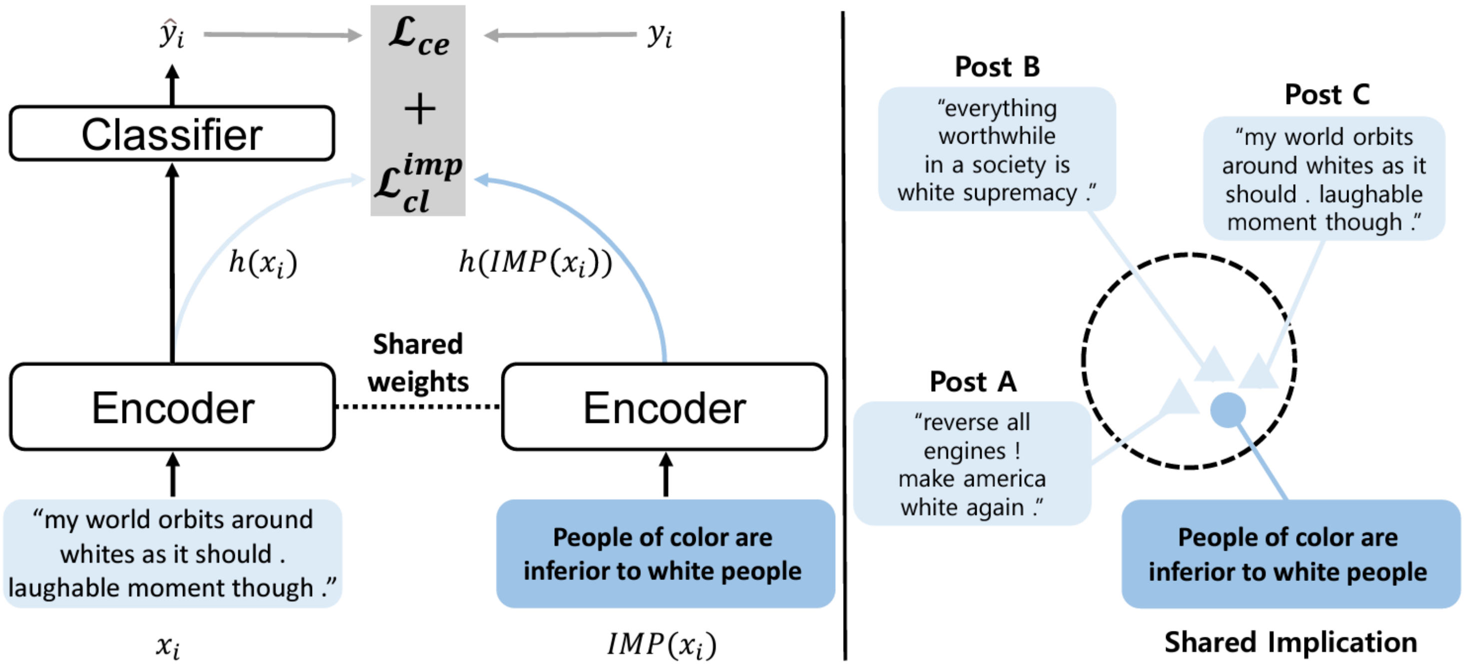
Generalizable Implicit Hate Speech Detection using Contrastive Learning

Youngwook Kim[1], **Shinwoo Park**[2] and **Yo-Sub Han**[1]
[1]Department of Computer Science, Yonsei University, Seoul, Republic of Korea
[2]Department of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea
{youngwook, pshkhh, emmous}@yonsei.ac.kr

Implication extraction

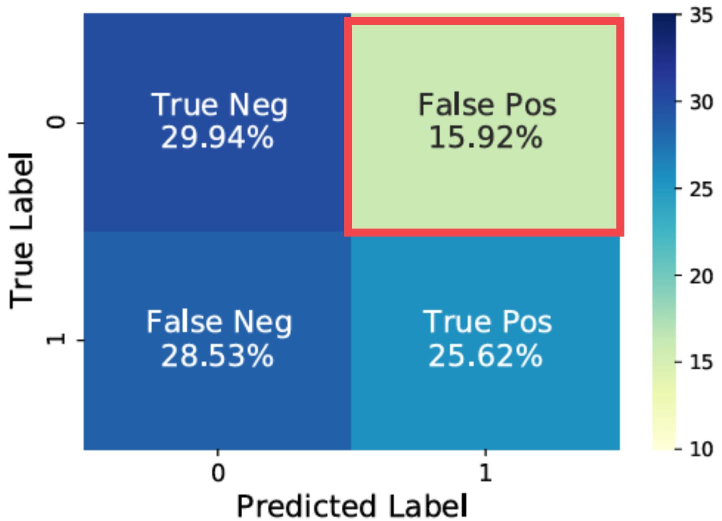
- Implication extraction (Kim et al 2022: Generalizable Implicit Hate Speech Detection using Contrastive Learning)





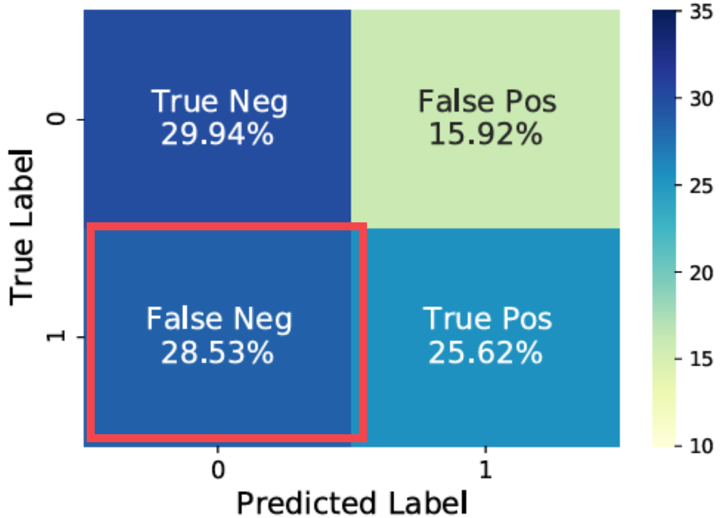
$\hat{y}_i$
$\mathcal{L}_{ce}$
+
$\mathcal{L}_{cl}^{imp}$
y_i
Classifier
$h(x_i)$
$h(IMP(x_i))$
Shared weights
Encoder
Encoder
"my world orbits around whites as it should . laughable moment though ."
People of color are inferior to white people
x_i
$IMP(x_i)$
Post B
"everything worthwhile in a society is white supremacy ."
Post C
"my world orbits around whites as it should . laughable moment though ."
Post A
"reverse all engines ! make america white again ."
People of color are inferior to white people
Shared Implication

Model	Objective	IHC → IHC (In-dataset)
BERT	CE	0.777
BERT (Aug)	CE	0.777
BERT	CE + *SCL*	0.777
BERT	CE + *AugCon*	0.774
BERT	CE + *ImpCon*	0.780
BERT	CE + *AugCon* + *ImpCon*	0.779
HateBERT	CE	0.764
HateBERT (Aug)	CE	0.763
HateBERT	CE + *SCL*	0.767
HateBERT	CE + *AugCon*	0.765
HateBERT	CE + *ImpCon*	0.774
HateBERT	CE + *AugCon* + *ImpCon*	0.772

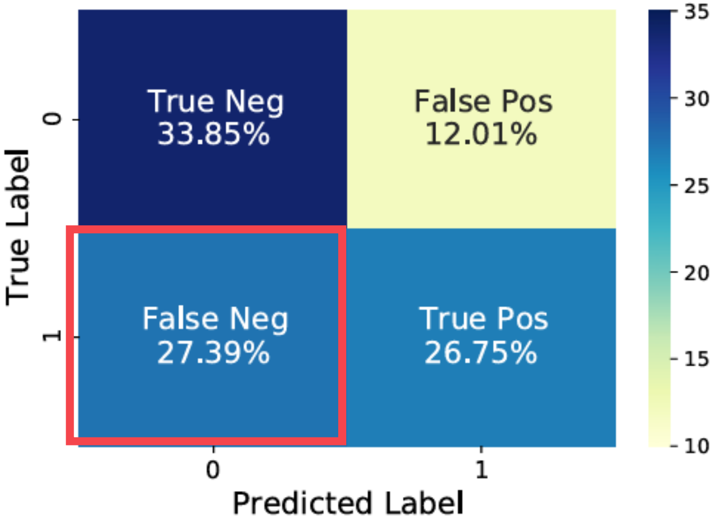


(a) CE (b) CE + *ImpCon*

- A target group that rarely appears in the training set



(a) CE



(b) CE + *ImpCon*

- A target group that rarely appears in the training set

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 3 September 2023, accepted 19 September 2023, date of publication 25 September 2023, date of current version 29 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318863

RESEARCH ARTICLE

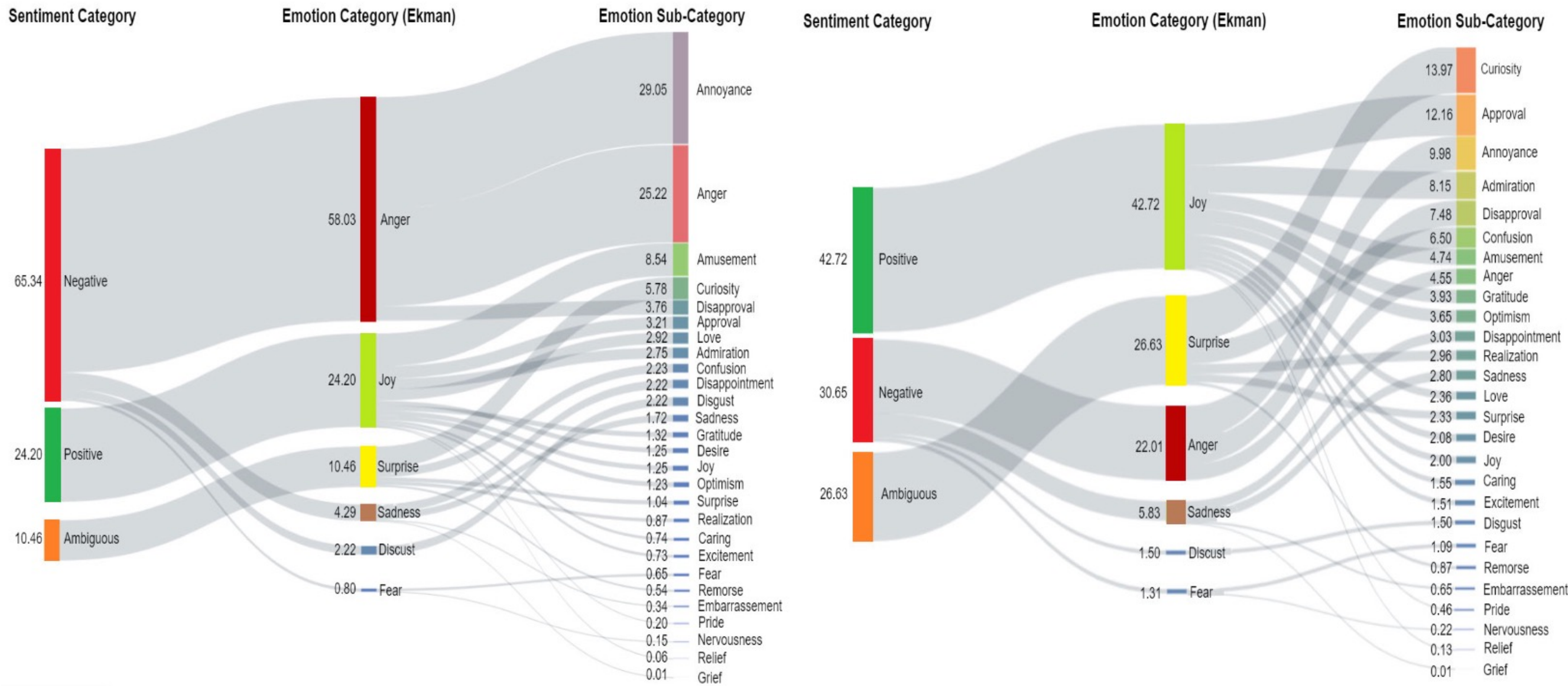
Fine-Grained Emotions Influence on Implicit Hate Speech Detection

AMIR REZA JAFARI, GUANLIN LI, PRABODA RAJAPAKSHA, (Member, IEEE), REZA FARAHBAKHSH, (Member, IEEE), AND NOEL CRESPI, (Senior Member, IEEE)
Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France

Corresponding author: Amir Reza Jafari (amir-reza.jafari_tehrani@telecom-sudparis.eu)

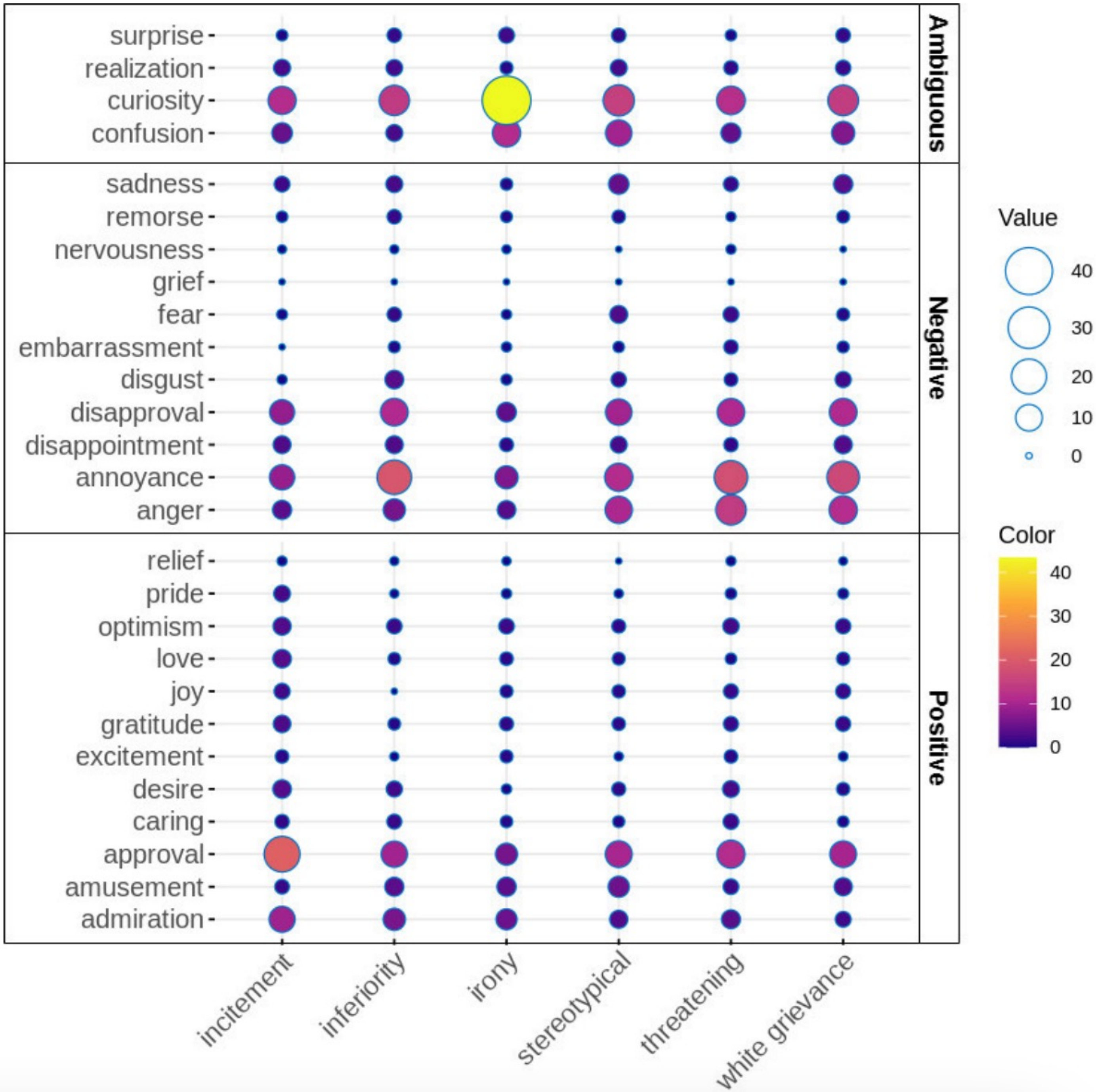
Contributions:

- An extensive analysis of emotion distribution of hateful vs non-hateful content
- Implicit hate speech classification with single-task learning (STL) and multi-task learning (MTL) models
- Plus a nice survey about implicit hate speech methods



(a) hateful content

(b) non-hateful content



LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Emotion Extraction

TABLE 3. Implicit hate categories definition proposed by [5] with examples and emotion model output.

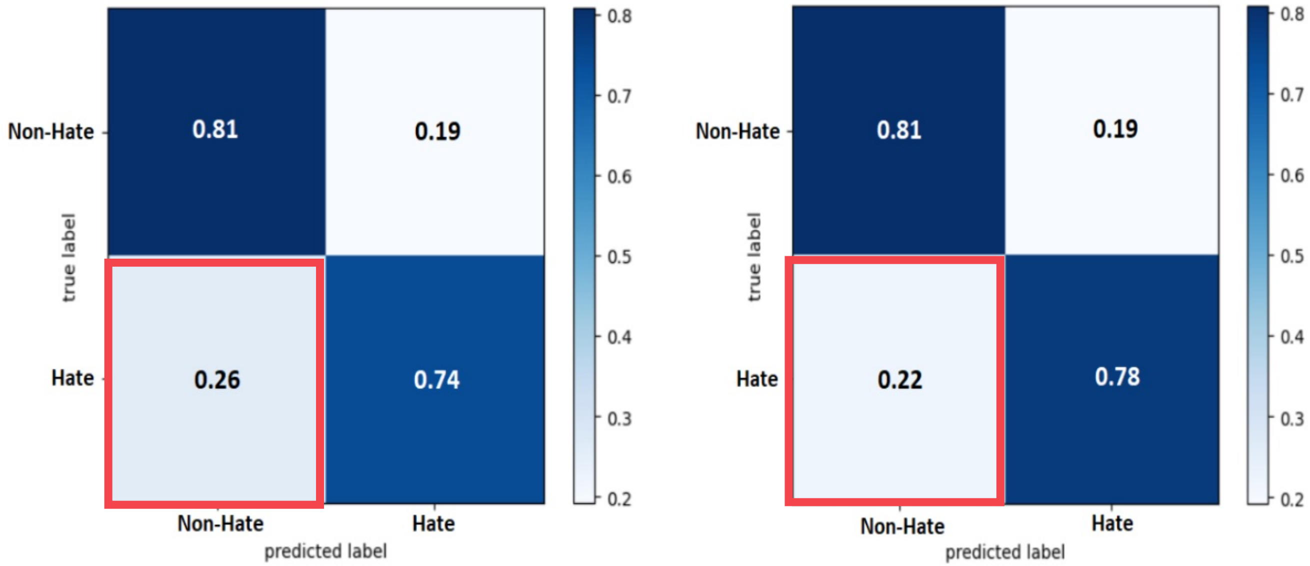
Category	Definition	Example tweet	Model output
Incitement	flaunting in-group unity to the implicit promotion of known hate groups and ideologies	this is an inherently white thing to do. keep up the good work	Admiration
Inferiority	using toxic language to show inferiority over a group	I really hate neocons like but not black people. it is not your fault that you are black	Anger
Irony	using sarcasm, humour, and satire to attack a group or individual	would you like fries with your mass amnesty and depressed wages?	Curiosity
Stereotypical	using negative attributes such as terrorism and metaphorical language to associate a group	You're Middle Eastern? Really? It's so good to know there are actually decent Middle Eastern people out there.	Admiration, Curiosity, Surprise
Threatening	attacking a group or individual with targeting pain, injury, damage, and violation	we need to stop the flow of immigration in our country! all must be vetted! just obey the laws! deport criminals!	Anger
White grievance	showing frustration over a minority group	not a good time to be an old white guy	Disapproval

TABLE 4. Experimental results of the STL models for the binary classification of implicit hate speech. F1 scores are reported in the macro average.

Feature level	TF-IDF				GloVe				BERT			
	Precision	Recall	macro-F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc
Text-only (Latent Hatred) [5]	59.5	68.8	63.9	71.6	56.5	65.3	60.6	69.0	72.1	66.0	68.9	**78.3**
Sentiment	63.6	67.3	64.4	71.5	59.0	67.6	63.0	70.7	72.4	73.5	72.8	75.4
Ekman level	63.6	69.0	66.2	72.4	59.0	67.4	62.9	70.6	72.2	73.6	72.9	76.4
Fine-grained Emotion	**64.7**	67.0	65.8	71.4	**60.5**	67.1	63.6	70.9	72.7	**74.3**	**73.5**	77.2
All features	64.4	**69.1**	**66.7**	**72.6**	60.3	**67.9**	**63.9**	**71.8**	**72.9**	74.0	73.4	75.9

Effect of MTL on accuracy

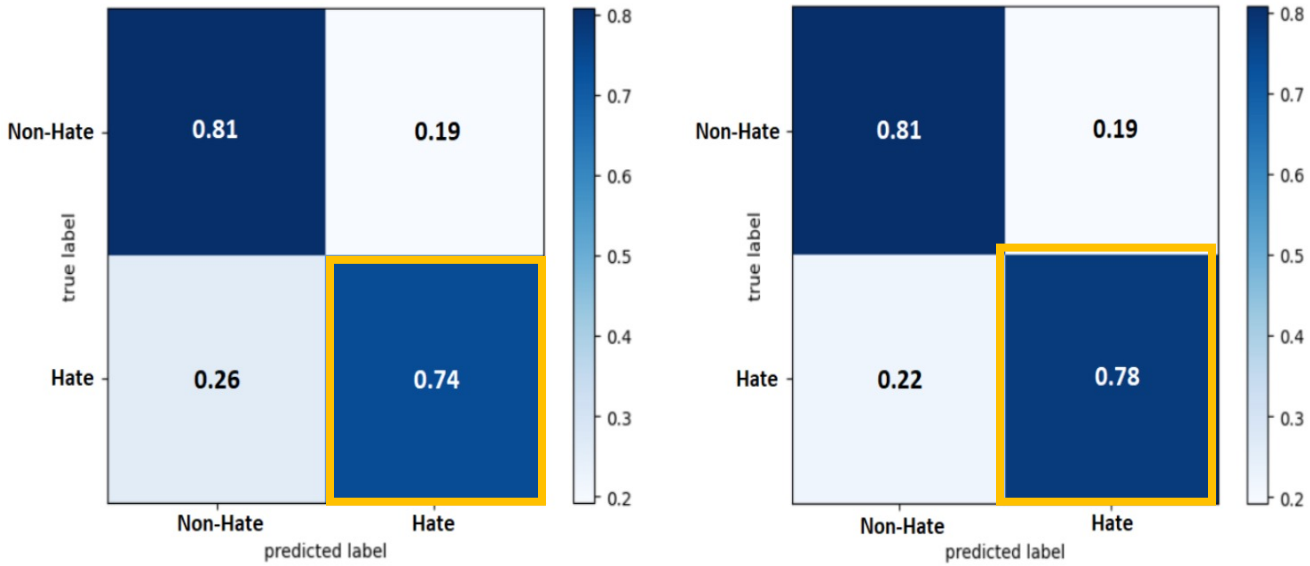
- Multi-task learning (with fine-grained emotion labeling) instead of single task learning



(a) Single-task learning (b) Multi-task learning

Effect of MTL on accuracy

- Multi-task learning (with fine-grained emotion labeling) instead of single task learning



(a) Single-task learning (b) Multi-task learning

Effect of MTL on implicit hate categories

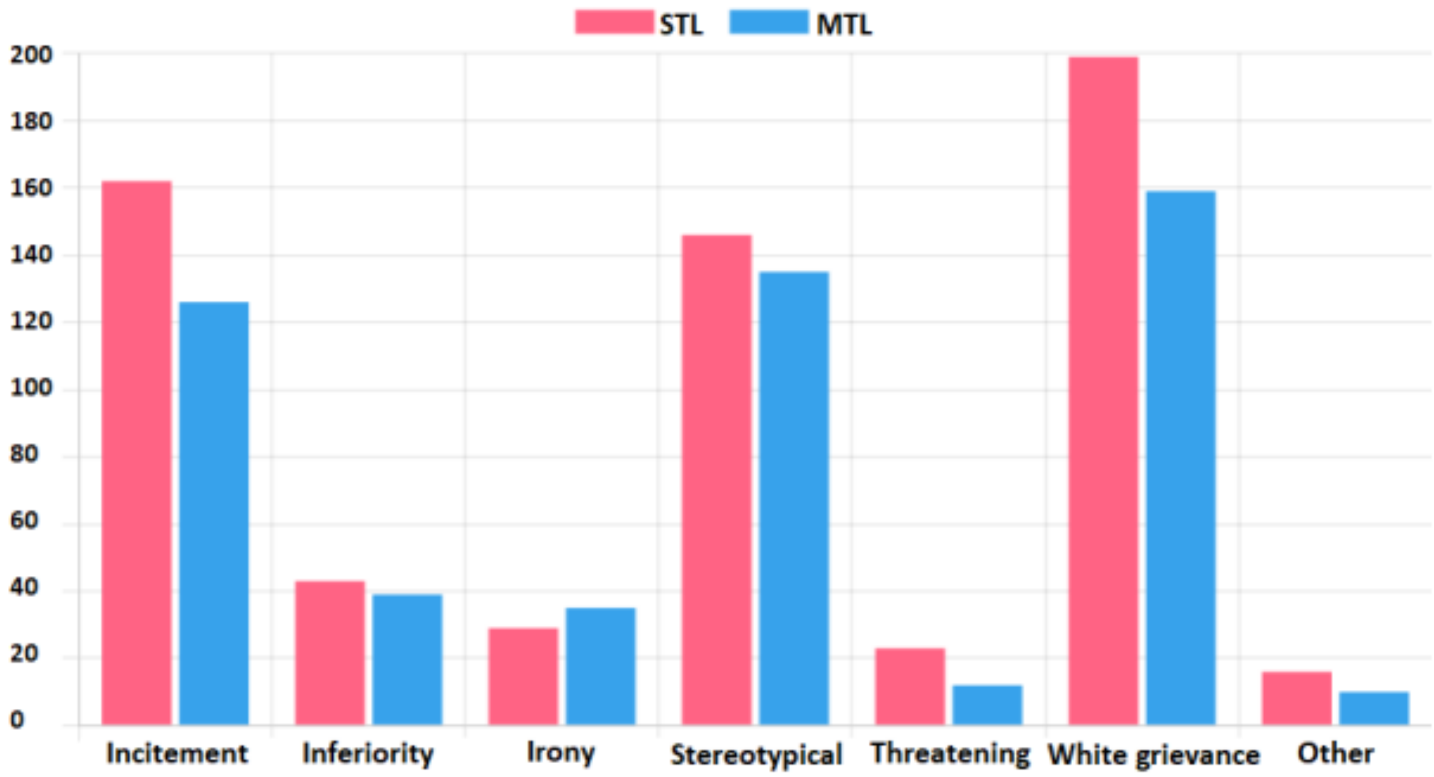


FIGURE 6. Comparing STL and MTL in number of false negatives for each implicit hate category test set.