Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization

Sarah Masud[1], Manjot Bedi[2], Mohammad Aflah Khan[1], Md Shad Akhtar[1], Tanmoy Chakraborty[1]
[1]IIIT-Delhi, [2]India & Northeastern University, USA
{sarahm, aflah20082, shad.akhtar, tanmoy}@iiitd.ac.in & bedi.m@northeastern.edu

LCS2 LABORATORY FOR COMPUTATIONAL SOCIAL SYSTEMS

Problem Motivation

Proactive mitigation of hate speech is an intervention step that is applied before the content is made public in the first place.

Hate intensity/severity (ϕ) of hate speech can be defined as the explicitness or hate, containing direct attacks, offensive lexicons and mentions of the target entity, among others.

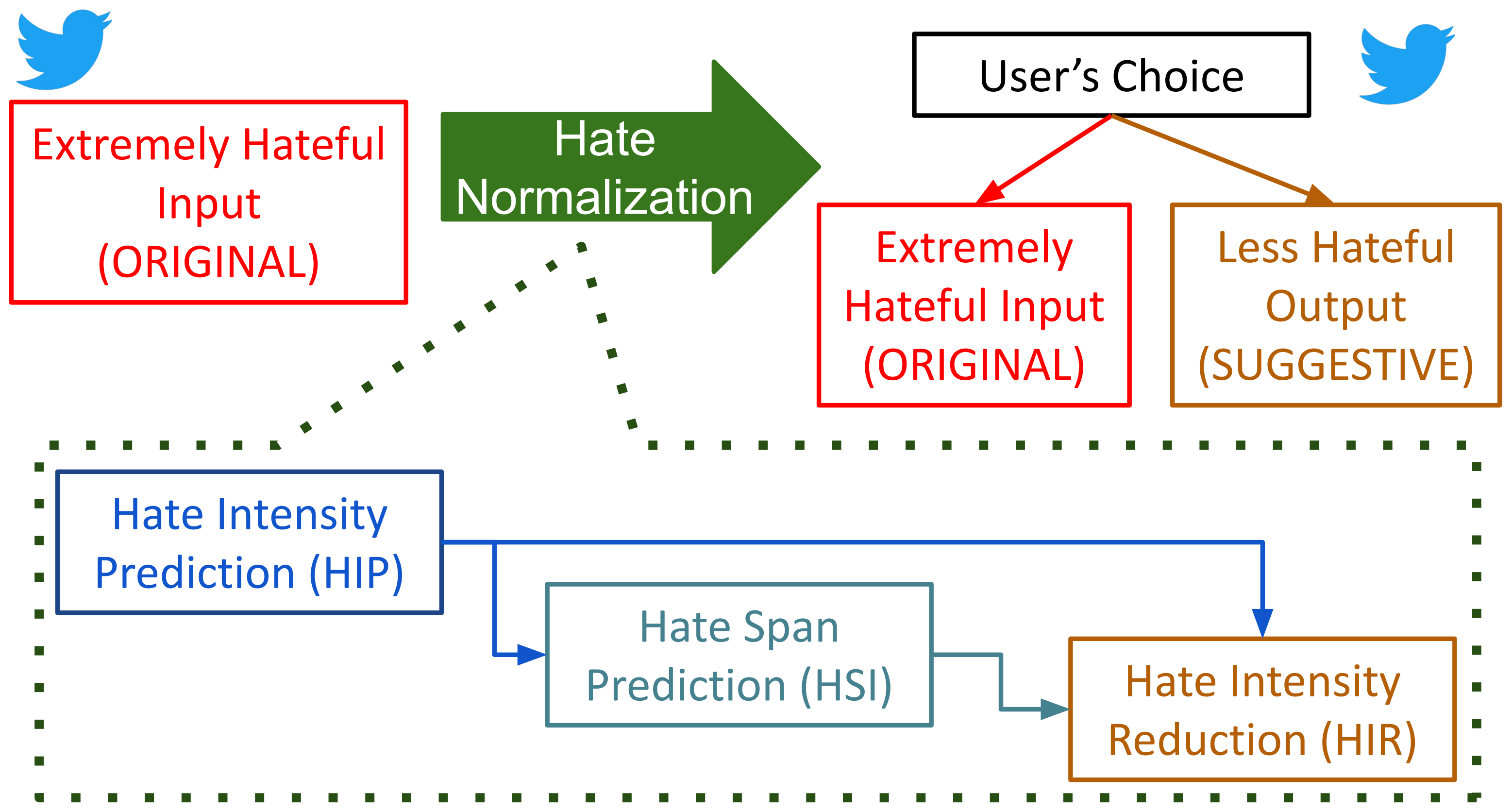
Hate normalization is the task of rephrasing a sentence with high hate intensity into a sentence with less hate intensity while still maintaining the hostile intent. In the example below, the hate intensity of original and normalised samples is 8 and 4 respectively.

Org	*This {immigrant should be hung or shot ! Period ! An***}$_{Span}$. @user*
Norm	This immigrant should be punished . @user

OBJECTIVE:

For a given high intensity hate sample t, our objective is to obtain its normalized form $t`$ such that the intensity of hatred $\phi_t$ is reduced while the meaning still conveys. $\phi_{t`} < \phi_t$

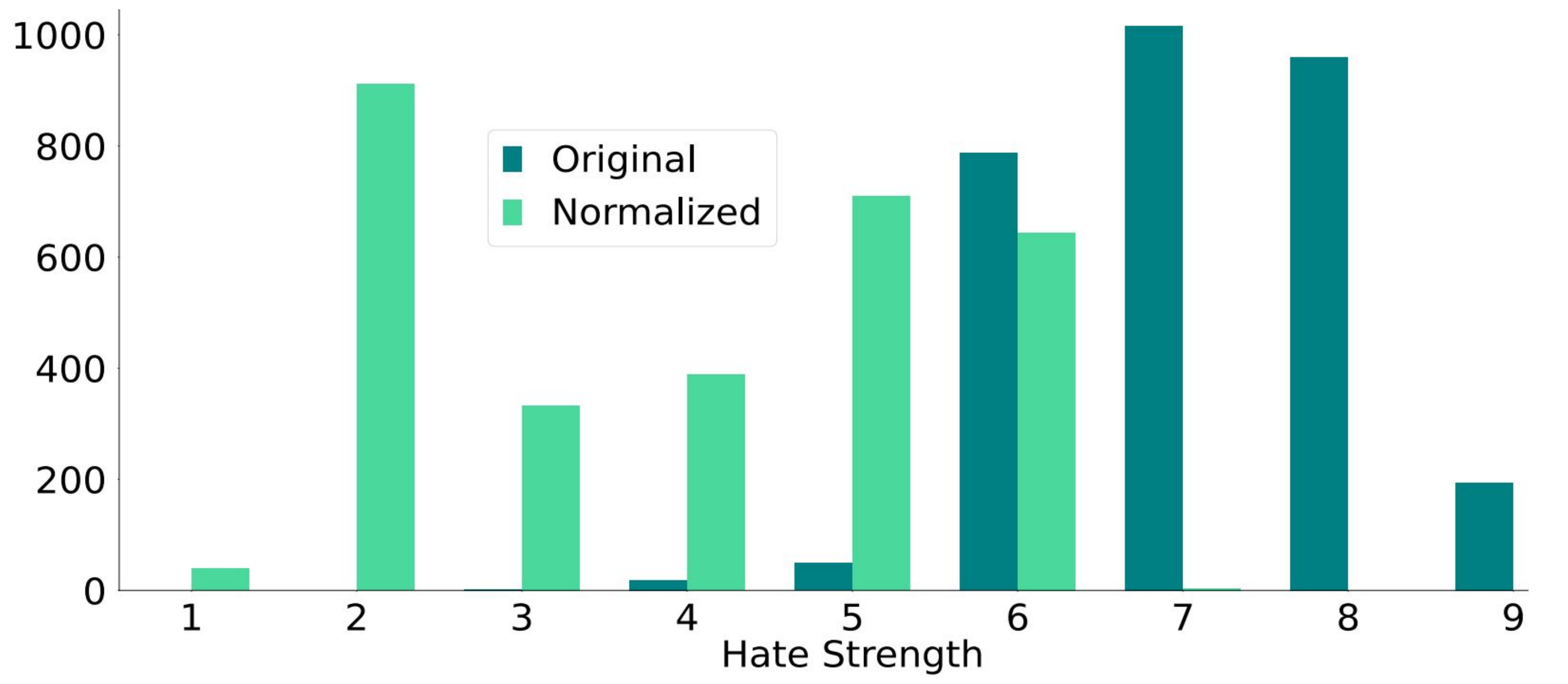
Workflow



Extremely Hateful Input (ORIGINAL)
Hate Normalization
User's Choice
Extremely Hateful Input (ORIGINAL)
Less Hateful Output (SUGGESTIVE)
Hate Intensity Prediction (HIP)
Hate Span Prediction (HSI)
Hate Intensity Reduction (HIR)

Dataset

- Hateful samples are manually annotated for intensity and hateful spans.
- Manual generation of normalised counter-part and its intensity.
- Hate intensity is marked on a scale of 1-10, 1 being lowest.
- Observable shift in hate intensities towards left.



References

[1]: Katsaros et al. *Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content*, ICWSM 2022.
[2]: Pryzant et al. *Automatically Neutralizing Subjective Bias in Text*, AAAI 2020.
[3]: Dai et al. *Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation*, ACL 2019.
[4]: Shen et al. Style Transfer from Non-Parallel Text by Cross-Alignment, NeurIPS 2017
[5]: Luo et al. *Towards Fine-grained Text Sentiment Transfer,* ACL. 2019

Proposed Methodology

Proposed method: **NACL** (**N**eural h**A**te spee**C**h norma**L**izer) is composed of:

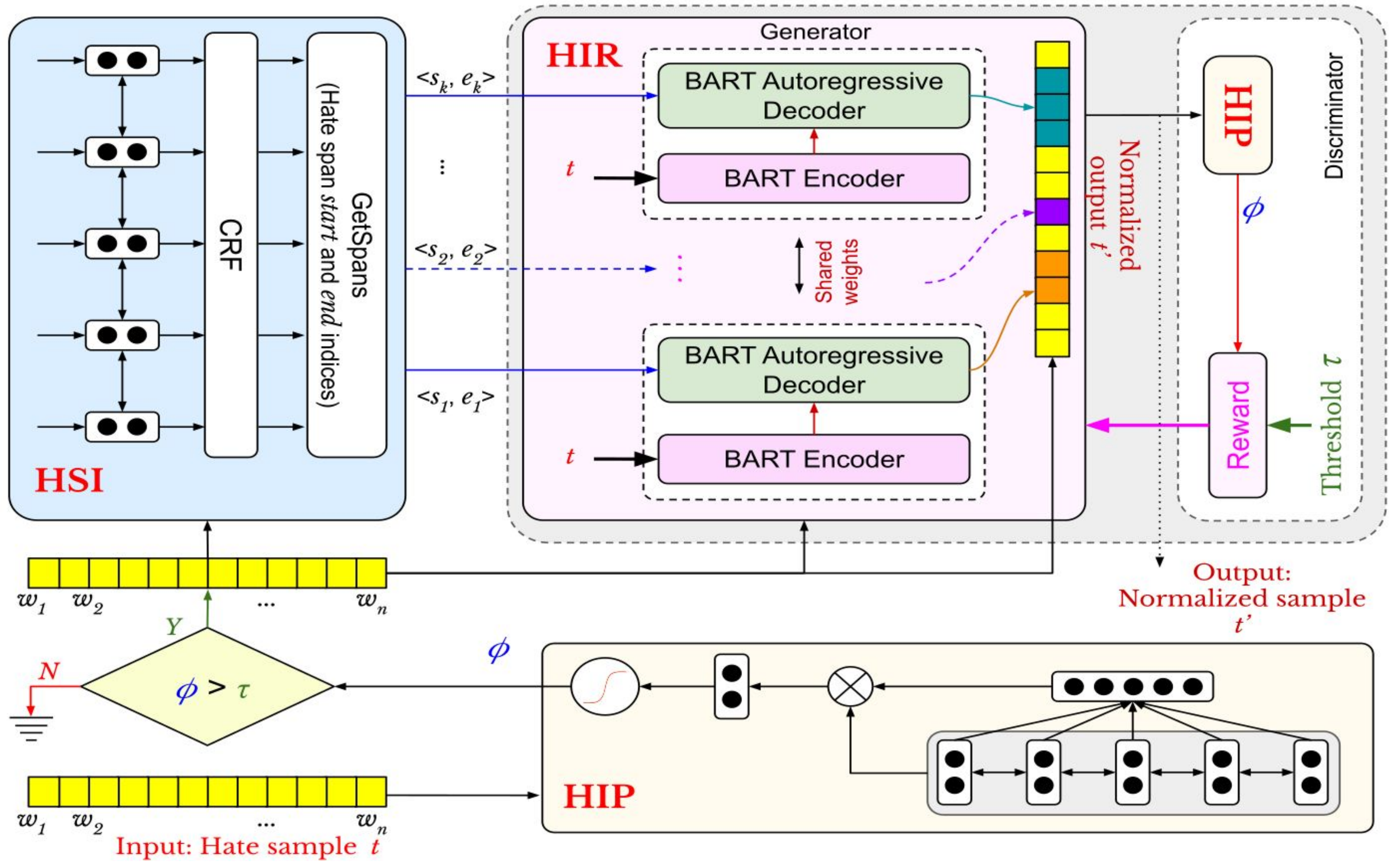
- **HIP** (Hate Intensity Prediction): HIP is a BiLSTM based regressor.
- **HSI** (Hate Span Identification): HSI is a BiLSTM + CRF predictor.
- **HIR** (Hate Intensity Reduction): BART based decoder with pre trained HIP as discriminator.

Reward

$$R_{t'} = \tau - \phi_{t'}$$

Overall Loss

$$L = l + (1 - R)$$



Results

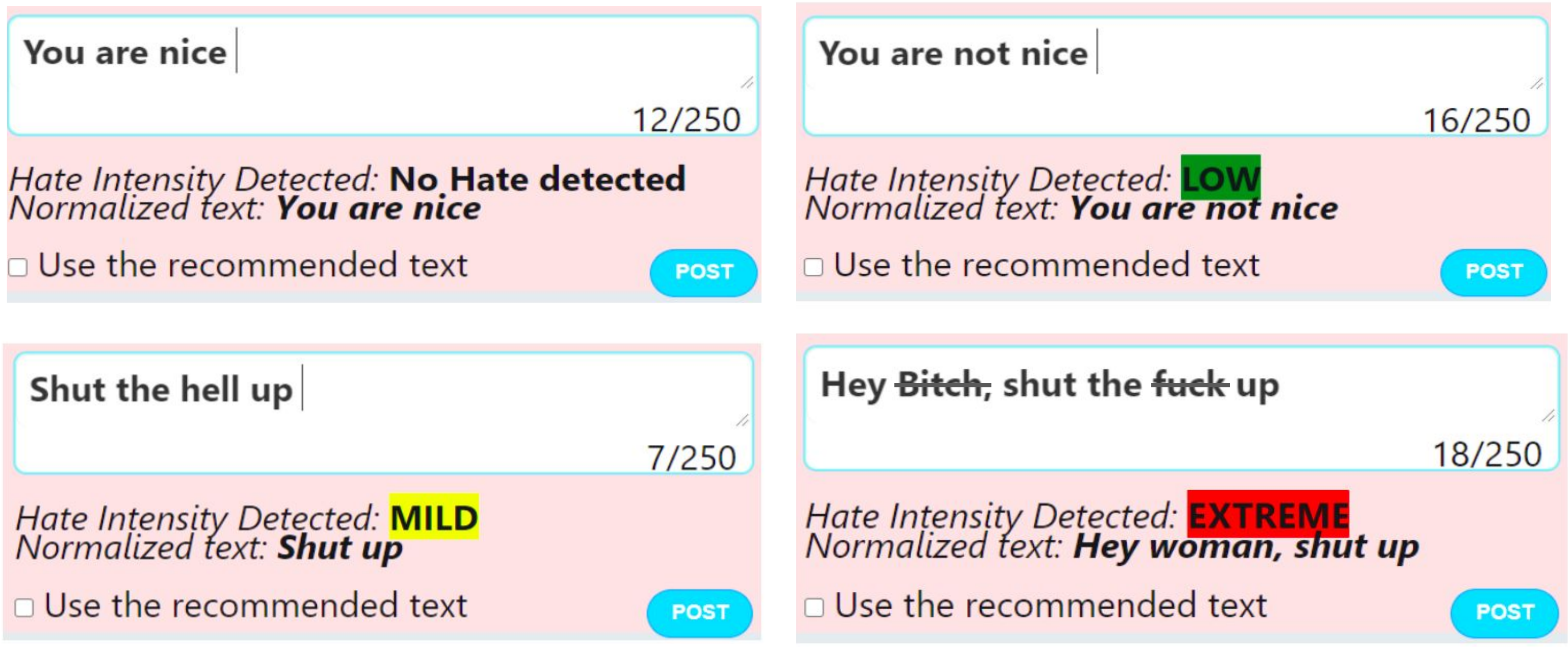
Supervised	Model	Evaluation Measure	
		BLEU ↑	Perplexity ↓
Yes	Dictionary Model	55.18	92
	Bias Neutralization	39.48	90.38
No	FGST	39.35	123.38
	Style Transformer (ST)	15.55	200.85
	Style Transfer (NPTCA)	0.93	1138.4
	Style Transfer (DRG)	0.84	199.58
Yes	NACL-HSR (τ=3)	58.84	86.11
	NACL-HSR (τ=5)	**82.27**	**80.05**
	Gold	100	64.66

Hate detection method	Normalization Model					
	FGST	Bias	ST	DRG	NPTCA	NACL-HIR
Waseem and Hovy [47]	0.00	0.03	0.03	−0.02	−0.04	0.03
Davidson et al. [11]	0.04	0.00	0.00	0.35	0.21	0.26
Founta et al. [19]	0.04	−0.01	0.07	0.23	0.04	0.03

- Comparison of our generation module NACL-HSR against baselines and gold standard. The proposed model with threshold intensity of five performs best.
- Performance of normalisation methods in reducing hate class confidence of existing hate classifiers.

Sample Output & Demo Tool

- Snapshots of the web extension for four scenarios. NACL generates normalized text only if $\phi_t > \tau$.
- The web framework detects hate as the user types in, and if any $\phi_t > \tau$ it shows the level of hate in the current text, and then recommends the normalized text to the user.



You are nice
12/250
Hate Intensity Detected: **No Hate detected**
Normalized text: ***You are nice***
□ Use the recommended text
POST

You are not nice
16/250
Hate Intensity Detected: **LOW**
Normalized text: ***You are not nice***
□ Use the recommended text
POST

Shut the hell up
7/250
Hate Intensity Detected: **MILD**
Normalized text: ***Shut up***
□ Use the recommended text
POST

Hey ~~Bitch,~~ shut the ~~fuck~~ up
18/250
Hate Intensity Detected: **EXTREME**
Normalized text: ***Hey woman, shut up***
□ Use the recommended text
POST

Enlisted below are few erroneous examples of NACL-HIR-generated vs. gold normalized texts

Type	Example
Original	*#LateNightThoughts how many Congressman {d***s did women s**k} to finally gain voting rights*
Reference	*#LateNightThoughts how many Congressman {did women approach} to finally gain voting rights*
Generated	*#LateNightThoughts how many Congressman {did women s**k} to finally gain voting rights*
Original	*{S**s are half breed trash}. No {filthy native} should be allowed to speak to any European.*
Reference	*No {native} should be allowed to speak to any European.*
Generated	*{Mexicans are t**h}. No {disgusting native} should be allowed to speak to any person.*

Conclusion

- We introduce a proactive measure of countering hate speech via normalization.
- In the current work, we skipped over the implicit hateful samples due to the absence of explicit hate spans.
- In future we wish to extend NACL to non-English texts.

Travel Support

acm-w celebrate support advocate
KDD 2022