EMNLP 2023

6-10 December, 2023

University of Maryland, College Park

Netaji Subhas University of Technology, New Delhi

CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network

SREYAN GHOSH*, MANAN SURI,* PURVA CHINIYA*, UTKARSH TYAGI*, SONAL KUMAR*, DINESH MANOCHA

* Equal Contribution

Table of Contents

		Page
I	Introduction	3
II	Methodology	7
III	Experiments and Results	14
IV	Conclusion	17

I Introduction: Implicit Hatespeech

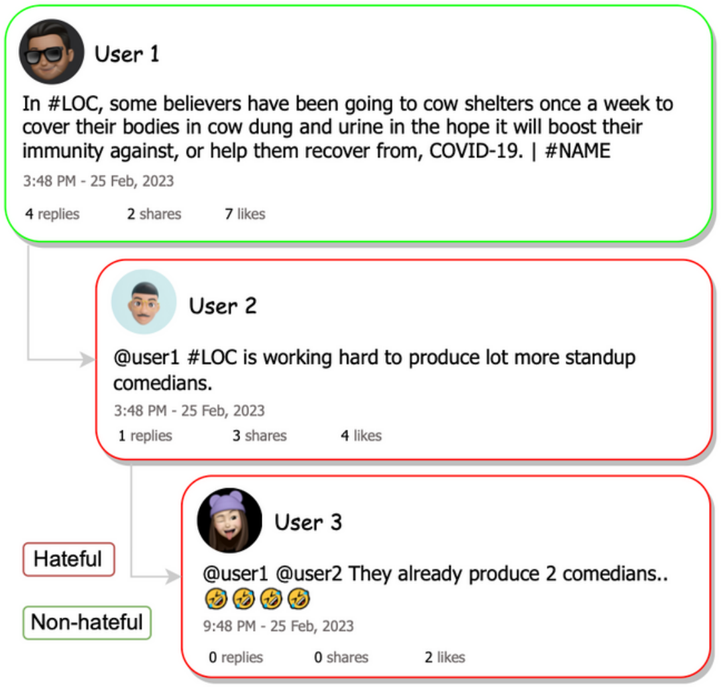


Figure 1: Illustration of a social media conversation tree with *implicit hate speech*. User 1 posts a factual statement about practices people follow in a certain location. In response, User 2 implies hate through a sarcastic statement, to which User 3 elaborates with a positive stance. Clearly, these utterances are even difficult for humans to classify as hate or not without the proper conversational context.

Hatespeech can take the form of overt abuse, i.e. **explicit hatespeech**, or it can be expressed in oded or indirect language i.e. **implicit hatespeech.**

I Introduction: Challenges in Implicit Hatespeech

While detecting explicit hatespeech is a popular task in NLP, by virtue of it being easier to detect because of overt linguistic signals, detecting implicit hatespeech has certain challenges.

1. **Linguistic nuance and diversity:** Implicit hate can be conveyed through sarcasm, humour, euphemisms, circumlocution, and other symbolic or metaphorical languages
2. **Varying Context:** Implicit hatespeech can be conveyed through everything, from dehumanising comparisons and stereotypes to threats, intimidation and incitement to violence.
3. **Lack of sufficient linguistic signals:** Unlike parent posts, which contain sufficient linguistic cues through background knowledge provided by the user, replies or comments to the parent post are mostly short and context-less reactions to the parent post, making implicit hate speech difficult to detect and emphasizing the need for better learning systems.

I Introduction: Limitations in Previous Work

DEFINITION AND INITIAL DATASETS

- ElSherief et al. (2021) define implicit hate speech as coded or indirect language disparaging individuals or groups.
- Latent Hatred, proposed by ElSherief et al., serves as a benchmark dataset for model performance in implicit hate speech classification.

LIMITATIONS OF LIN'S SYSTEM

- Lin (2022) extends Latent Hatred by leveraging knowledge graphs (KGs) for implicit hate speech classification.
- Restriction to English and unavailability of KGs in other languages limit the system's applicability.
- Lin's system also fails to capture external context, a crucial element for effective hate speech detection (Sheth et al., 2022).

SHORTCOMINGS OF LATENT HATRED AND TAXONOMY

- Existing state-of-the-art classifiers struggle to perform well on Latent Hatred, indicating limitations in current approaches.
- Latent Hatred's 6-class taxonomy overlooks conversational-context-sensitive implicit hate speech, a significant portion online (Modha et al., 2022; Hebert et al., 2022).

PROPOSAL FOR IMPROVEMENT

- The need to extend the definition of implicit hate speech to include hate conveyed in the context of conversational dialogue.
- Introduction of a novel neural learning system to address the limitations of prior approaches.
- Emphasis on the importance of capturing external context for more effective and inclusive hate speech detection.

I Introduction: Major Contributions

- We introduce **CoSyn**, the first neural network architecture specifically built to detect implicit hate speech in online conversations. CoSyn leverages the strengths of existing research and introduces novel modules to **explicitly take into account user and conversational context integral to detecting implicit hate speech**.

- We provide **implicit hate speech annotations** for 6 popular hate speech datasets.

- Through extensive experimentation, we show that CoSyn outperforms all our baselines quantitatively on **6 hate speech datasets** with **absolute improvements of 1.24% - 57.8%**.

- We also perform **extensive ablative experiments** and **qualitative comparisons** to prove the efficacy of CoSyn.

II Methodology

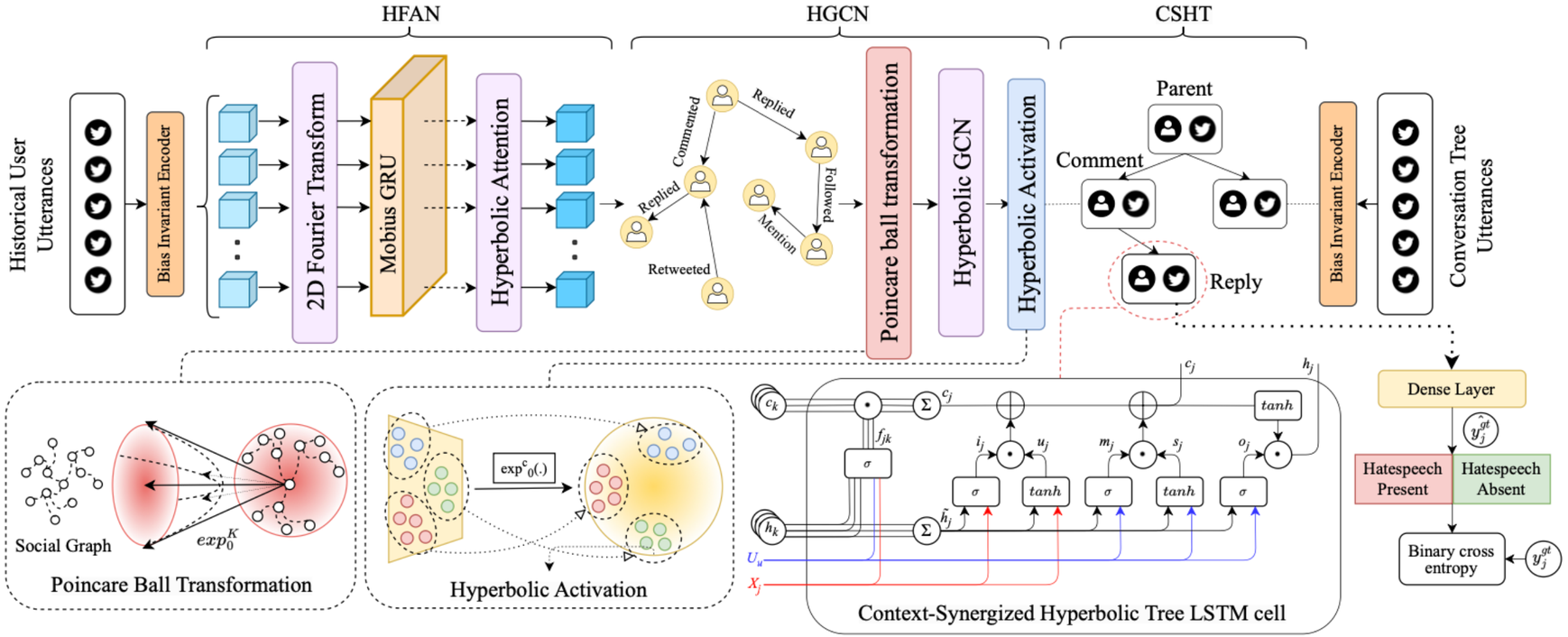
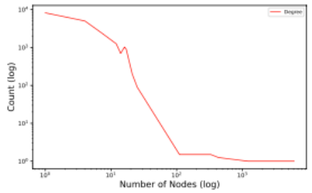


Figure 3: Illustration of **CoSyn**. CoSyn detects if a target utterance, authored by a social media user, implies hate or not leveraging three main components: (1) **HFAN**: This models the user's personal historical context by employing Fourier Transform and Hyperbolic Attention on the user's historical tweets. (2) **HGCN**: This models the user's social context through hyperbolic graph learning on its relations with other users. (3) **CSHT**: This jointly models the user's personal context and the conversational context to finally classify if the utterance is hateful.

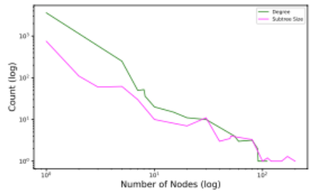
II Methodology: Motivation and Design for Different Components

Component	Motivation	Design
Bias Invariant Encoder	To learn bias-invariant representations, specifically, handle keyword bias	Fine-tune a SentenceBERT and solve an additional loss proposed by Mathew et al., 2021 using self-attention maps and ground-truth hate spans.
Hyperbolic Fourier Attention Network	To learn the user's *personal historical context*.	HFAN facilitates varied user engagement on social media by employing Discrete Fourier Transform for abstract frequency modeling, a Möbius GRU for temporal distribution modeling, and hyperbolic attention on prior user utterances.
Hyperbolic Graph Convolutional Network	To learn the user's *personal social context*.	HGCN employs hyperbolic learning to capture social network dynamics by utilizing user connections as edges in the graph.
Context-Synergized Hyperbolic Tree-LSTM	To jointly model a user's *personal context* and the *conversational dialogue context*.	CSHT efficiently models scale-free conversation trees, capturing context interactions within a hyperbolic learning framework. It modifies a tree LSTM backbone to work in hyperbolic space, and account for both the utterance and user context.

II Methodology: Modelling Scale Free Properties in Social Media



(a) Frequency distribution for the social graph.



(b) Frequency distribution for conversation trees.

Property	Social Graph	Conversation Tree
Mean Degree	18.22	2.2
Node Degree	2.66E-03	2.89E-04
δ	1.5	0.3
Power Law	$P(x) \sim x^{-\gamma}$	
γ	2.81	2.39

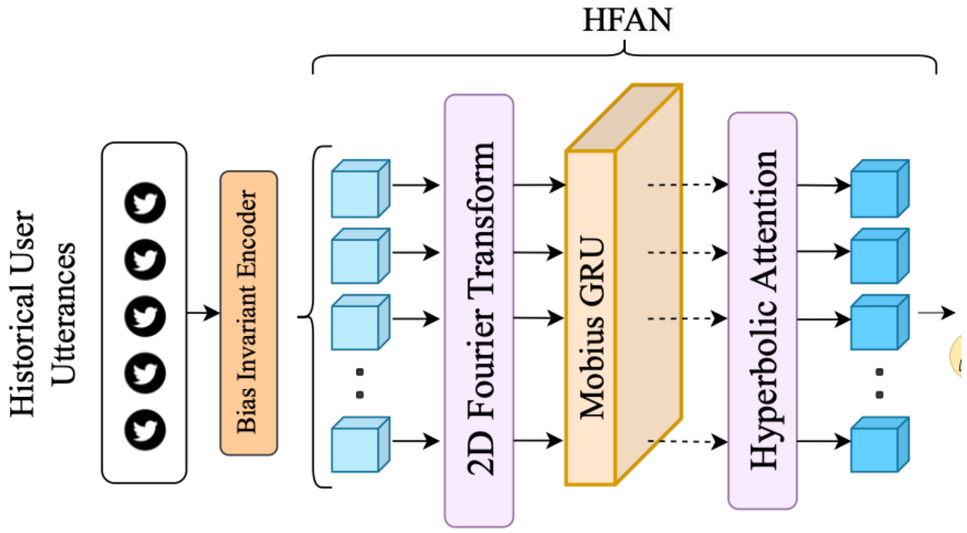
(c) Properties of the social graph and conversation trees.

Figure 2: Various properties of the social graph and conversation trees averaged across datasets. The fitting of the node distribution in the power law ($\gamma \in [2, 3]$) (Choromański et al., 2013) and low hyperbolicity (Barabási and Bonabeau, 2003) δ indicates the scale-free nature of conversations and social graphs.

- Social network connections between users on a platform often possess hierarchical and scale-free structural properties (degree distribution of nodes follows the power law as seen in Fig 2 and decreases exponentially with a few nodes having a large number of connections).

- Conversation trees on social media possess a hierarchical structure of message propagation, where certain nodes may have many replies; e.g., nodes that include utterances from popular users. Such phenomena lead to the formation of hubs within the conversation tree, indicating scale-free and asymmetrical properties of the conversation tree.

II Methodology: Modelling model a user's personal historical contex

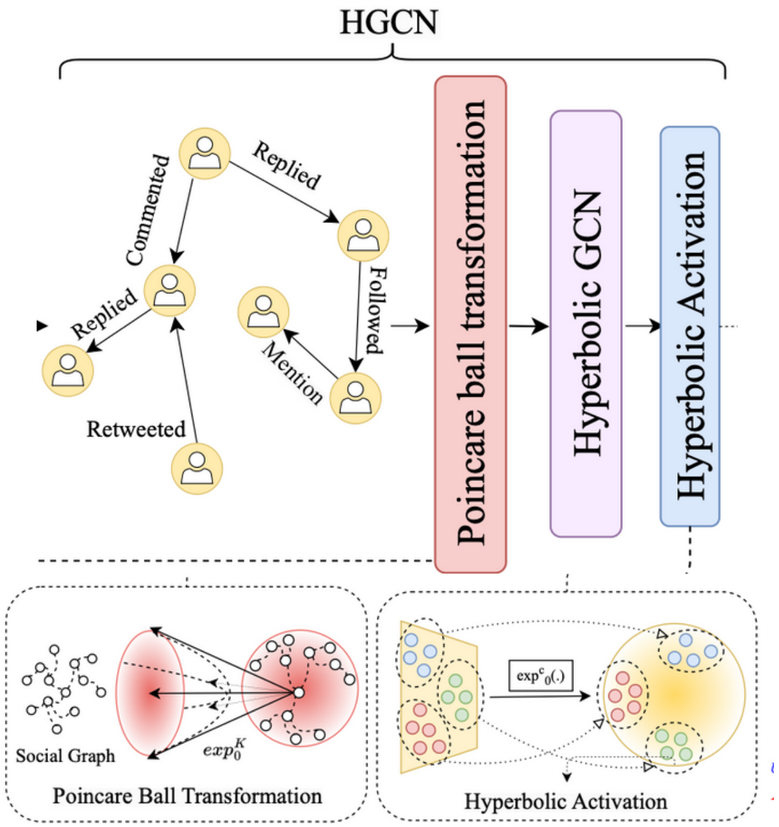


- Historical utterances of the user are encoded using our bias invariant encoder. These representations are processed using a 2D Fourier transform (along the temporal dimension and the embedding dimension. The 2D DFT operation helps highlight the most prominent frequencies, signifying a holistic understanding of the user's sociological behavior.

- Additionally, these frequencies may change over time. Thus, to account for the latter factor first, we pass the embeddings through a Hyperbolic GRU.

- Hyperbolic attention is used to find a distilled representation for each user which takes into context the different abstract frequencies and their temporal distribution in a scale-free manner.

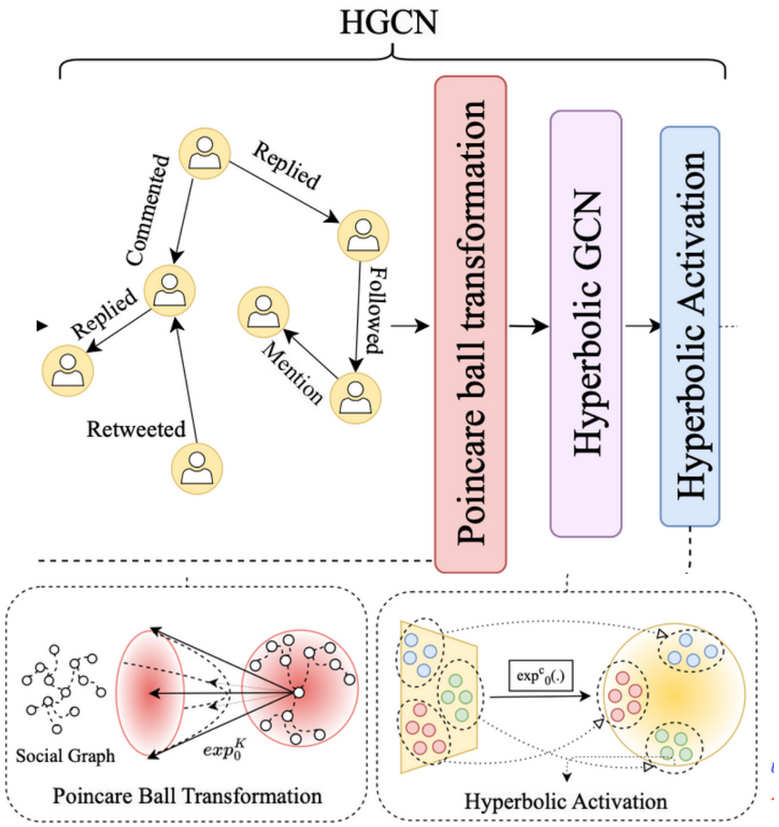
II Methodology: Modelling a user's personal social contex



- HGCN modifies the conventional GCN and performs neighbor aggregation using graph convolutions in the hyperbolic space to enrich a user's historical context representations learned through HFAN using social context.

- To model complex hierarchical representations, in social graphs, HGCN performs all operations in the Poincaré space.

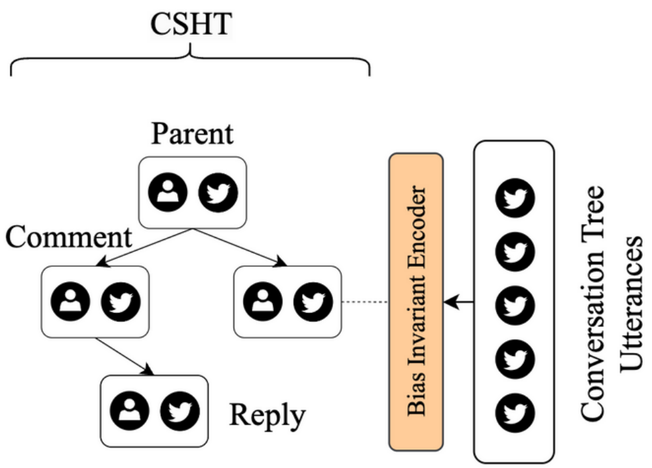
II Methodology: Modelling a user's personal social contex

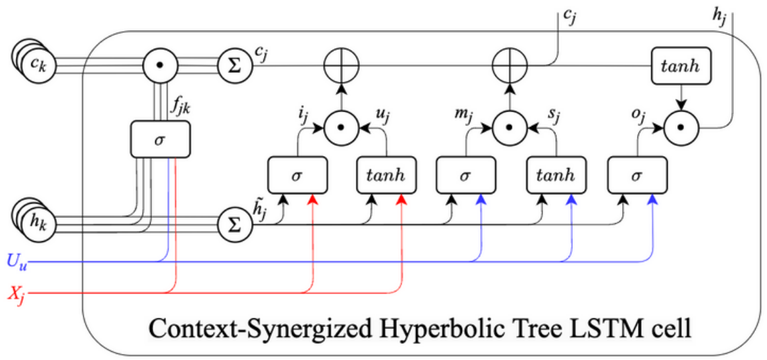


- HGCN modifies the conventional GCN and performs neighbor aggregation using graph convolutions in the hyperbolic space to enrich a user's historical context representations learned through HFAN using social context.

- To model complex hierarchical representations, in social graphs, HGCN performs all operations in the Poincaré space.

II Methodology: Jointly modelling the Conversational and User Context





- To model the conversational context in conversation trees effectively, we propose Context-Synergized Hyperbolic Tree-LSTM (CSHT). CSHT presents several modifications and improvements over Tree-LSTM, including:

(1) incorporating both the user's personal context and the conversation context while clearly capturing the interactions between them

(2) operating in the hyperbolic space, unlike the original TreeLSTM, which operates in the Euclidean space.

III Experiments and Results: Baseline Comparison

	Overall			Implicit					Overall			Implicit					Overall			Implicit				
	F_1	P	R	F_1	P	R	C. F_1	R. F_1	F_1	P	R	F_1	P	R	C. F_1	R. F_1	F_1	P	R	F_1	P	R	C. F_1	R. F_1
Baseline	**Reddit**								**CAD**								**DIALOCONAN**							
SentenceBert	71.12	63.35	**81.06**	76.05	77.01	75.11	22.45	18.06	46.89	51.30	43.18	49.01	48.03	50.03	35.40	–	46.23	37.20	**60.70**	34.75	26.46	50.60	**49.98**	29.55
ConceptNet	58.25	58.72	58.73	26.43	32.57	22.23	33.34	22.17	56.68	54.22	59.37	39.42	42.5	36.75	28.72	–	39.45	34.22	46.56	24.53	25.9	23.29	35.15	15.72
HASOC	56.21	51.86	61.35	28.9	31.48	26.71	35.50	20.35	50.11	47.54	52.97	40.61	46.75	35.89	30.29	–	43.67	32.45	66.74	27.22	23.80	31.78	28.45	17.02
Conc-Perspective	56.21	51.86	61.35	25.75	29.65	22.75	27.25	24.50	40.11	42.50	37.97	33.34	35.60	31.34	28.89	–	35.33	32.16	39.19	22.33	20.65	24.30	31.20	18.92
CSI	52.77	50.90	54.80	33.40	30.25	38. 28	34.00	20.50	65.80	63.18	68.64	43.14	44.25	42.00	30.35	–	42.23	45.01	39.77	26.25	28.90	24.05	30.60	17.55
GCAN	54.75	57.94	51.89	23.25	30.00	18.98	13.13	15.50	69.31	**71.00**	67.69	44.34	45.60	43.13	33.78	–	31.33	31.10	32.67	31.33	32.80	29.98	34.70	18.60
HYPHEN	53.44	55.12	54.38	25.00	27.10	23.20	27.40	20.00	42.65	40.11	45.50	32.80	30.22	35.86	29.98	–	34.67	38.80	31.34	16.40	20.00	13.89	19.19	12.30
FinerFact	63.25	62.14	64.40	27.11	30.25	24.56	40.55	25.70	65.36	64.00	66.77	26.25	30.12	23.26	18.22	–	30.70	28.89	32.70	15.60	18.20	13.60	21.15	18.11
Graph NLI	41.04	57.12	32.02	26.03	42.50	19.76	55.01	45.00	28.25	68.10	17.82	47.40	45.66	49.27	22.75	–	32.35	31.44	33.31	24.11	21.89	26.83	39.02	16.50
DUCK	60.10	61.89	58.40	34.23	26.26	49.15	40.55	32.30	30.66	32.50	29.02	28.11	24.25	33.43	22.15	–	28.60	28.20	29.01	20.45	18.23	23.28	26.89	14.40
CoSyn *(ours)*	**76.23**	**79.07**	73.58	**81.12**	**80.15**	**82.12**	**60.23**	**59.12**	**73.26**	70.07	**76.75**	**57.59**	**59.27**	**56.01**	**38.19**	–	**51.02**	**52.92**	49.25	**52.98**	**54.67**	**51.39**	48.77	**54.00**
	GAB								**ICHCL**								**Latent Hatred**							
SentenceBert	50.31	42.67	61.28	40.03	**49.78**	33.47	24.06	13.10	79.86	81.03	78.82	37.32	36.33	38.11	36.69	37.11	58.82	45.45	**83.33**	38.46	31.25	50.00	27.05	–
ConceptNet	47.82	48.72	46.95	19.29	25.23	15.61	16.12	10.12	69.11	68.23	70.01	28.29	28.27	28.31	28.91	27.89	48.12	46.87	49.43	37.23	34.45	40.49	22.56	–
HASOC	39.45	43.44	36.13	16.54	22.11	13.21	15.19	12.71	72.53	72.67	72.51	34.31	35.09	33.56	35.28	32.11	50.47	52.22	48.83	39.82	37.22	42.81	35.21	–
Conc-Perspective	51.47	47.23	56.54	18.23	20.54	16.38	24.29	18.45	71.18	70.14	72.25	31.18	29.59	32.94	30.46	29.82	51.22	53.79	48.88	40.11	38.12	42.31	34.37	–
CSI	49.62	51.22	47.17	20.60	22.24	19.18	23.50	18.34	69.47	75.20	64.55	26.50	24.01	30.15	23.64	19.31	56.25	53.11	59.70	42.06	32.80	58.59	21.25	–
GCAN	24.00	32.00	27.00	26.47	30.25	23.53	17.20	15.36	74.22	72.11	76.45	35.22	36.20	34.29	32.14	26.47	56.80	54.55	59.24	40.24	31.65	55.23	20.71	–
HYPHEN	48.32	45.19	51.92	23.50	25.21	22.00	28.60	19.45	72.72	67.11	74.41	33.34	27.66	41.95	34.66	34.21	53.20	51.6	54.9	42.40	**47.89**	38.04	34.11	–
FinerFact	53.00	52.65	53.35	18.50	24.00	15.05	23.68	15.04	69.11	67.43	70.87	32.27	33.82	30.85	24.90	18.65	52.11	48.32	56.54	52.04	42.67	**66.68**	32.00	–
Graph NLI	50.00	53.22	47.15	42.12	45.00	39.58	**47.00**	32.00	51.17	42.98	63.22	26.53	21.29	35.19	27.65	28.15	33.10	25.24	48.07	48.25	39.24	62.63	33.00	–
DUCK	62.78	60.50	65.30	35.70	36.85	34.62	36.20	24.60	78.36	78.42	78.30	37.88	37.64	38.12	36.48	35.59	56.00	58.50	53.75	35.15	30.00	42.44	26.16	–
CoSyn *(ours)*	**66.71**	**64.43**	**69.15**	**45.00**	37.01	**57.38**	46.22	**38.29**	**89.53**	**90.55**	**88.53**	**46.03**	**46.26**	**45.82**	**46.85**	**45.89**	**64.65**	**61.92**	67.63	**53.28**	47.66	55.49	**40.12**	–

Table 1: Result comparison of CoSyn with our baselines on 6 hate speech datasets. We compare performance on both the overall dataset and the implicit subset. C.F_1 and R.F_1 indicate F_1 scores measured on only comments and replies, respectively. CoSyn outperforms all our baselines with absolute improvements in the range of 3.4% - 45.0% when evaluated on the entire dataset and 1.2% - 57.9% when evaluated on only the implicit subset. – indicates conversation trees in the dataset did not have replies.

III Experiments and Results: Ablation Study

Ablations	Overall			Implicit				
	F_1	P	R	F_1	P	R	Comment F_1	Reply F_1
CoSyn (ours)	**70.23**	**69.83**	**70.82**	**56.00**	**54.17**	**58.04**	**46.73**	**49.32**
- DFT	67.54	68.24	69.29	54.52	52.57	57.34	45.22	46.92
- HFAN	66.62	65.32	66.23	54.68	51.78	58.04	45.44	47.04
- HGCN	66.56	66.98	65.92	53.28	52.02	56.98	45.12	46.32
- HFAN - HGCN	65.29	64.71	65.55	52.14	51.12	56.38	42.91	46.29
- User Context	62.31	62.94	64.21	48.72	48.94	49.53	39.88	41.19
BiCHST → UniCHST	68.67	68.23	68.36	55.29	52.78	57.89	46.09	47.53
Hyperbolic → Euclidean	66.47	66.48	67.21	54.83	54.14	56.41	45.44	47.41

Table 2: Ablation study on Cosyn. Results are averaged across all 6 datasets.

1. CoSyn's performance drops significantly without user context, emphasizing its importance. Using user context through HFAN and HGCN surpasses mere historical utterance embeddings in CSHT.
2. Modeling in Euclidean space results in a 3.8% overall F1 drop, reinforcing the effectiveness of hyperbolic space modeling.

III Experiments and Results: Qualitative Analysis

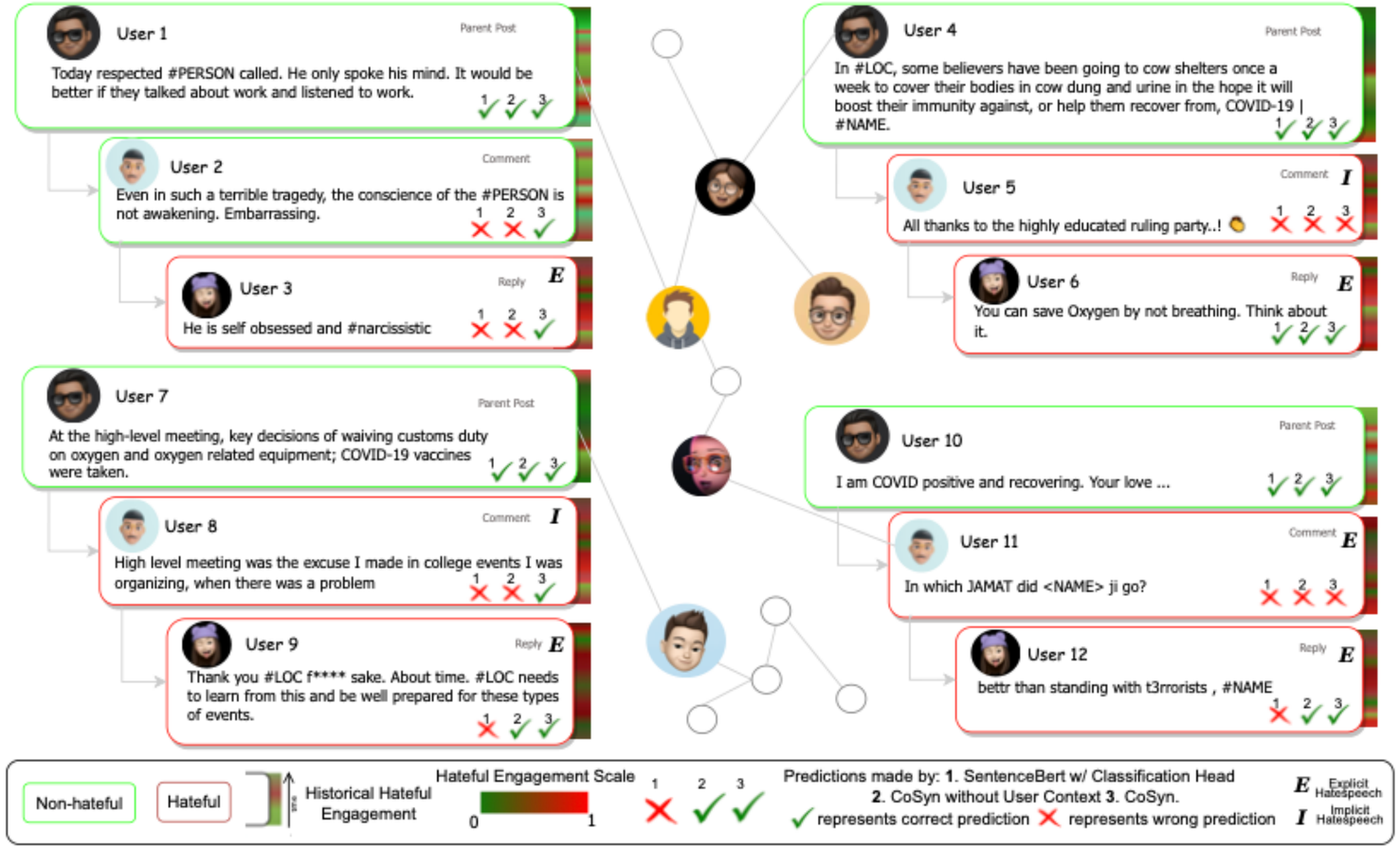


Figure 4: We study 4 conversation trees in the ICHCL dataset, including the prediction of different classifiers on the utterance to be assessed, the historical engagement of the author of the utterance, and the social relations between the different authors.

IV Conclusion

In this paper, we present CoSyn, a novel learning framework to detect implicit hate speech in online conversations. CoSyn jointly models the conversational context and the author's historical and social context in the hyperbolic space to classify whether a target utterance is hateful. Leveraging these contexts allows CoSyn to effectively detect implicit hate speech ubiquitous in online social media.

It is also important to acknowledge limitations of our work, which will inspire our future efforts.

Limitations

1. CoSyn's potential limitation lies in lacking world knowledge. Including this could significantly enhance its performance in this task (Sheth et al., 2022), a focus for future exploration.
2. Table 2 highlights that CoSyn's effectiveness depends on the seamless integration of its components. Future research will concentrate on boosting the performance of individual elements.

Thank you for listening!