

GYAAN-SAHIT: A Persona-Driven Multi-Agent Framework for Caste-Based Hate Speech Detection

Sakshi Gupta¹, Shunmuga Priya Muthusamy Chinnan², Saranya Rajiakodi³,
Ratnavel Rajalakshmi⁴, Bharathi Raja Chakravarthi²

¹School of Computer Science, University of Galway, Ireland,

²Data Science Institute, University of Galway, Ireland,

³Central University of Tamil Nadu, India,

⁴Vellore Institute of Technology, Chennai, India

Abstract

Social media has amplified public discourse in India while perpetuating caste-based hierarchies. Despite legal protections, caste-based hate speech continues to propagate across digital platforms through culturally embedded expressions that conventional classifiers often struggle to interpret. We propose GYAAN-SAHIT, a knowledge-driven multi-agent framework that addresses this problem through structured debate-based classification. Each agent adopts a distinct ideological and socio-cultural persona, engaging in multi-turn argumentation to reason over context, subtext, and intent. A critic agent then evaluates the coherence of the debate before producing the final classification. The framework further integrates Hindi hate lexicons to ground its reasoning in linguistic and cultural specificity. Experiments show that GYAAN-SAHIT achieves improvement in performance while generating culturally grounded explanations, demonstrating the effectiveness of persona-based multi-agent reasoning for hate speech detection in low-resource and socially complex environments.

Content Warning: This article contains illustrative examples of hateful content.

1 Introduction

The caste system in India is a long-standing social hierarchy that classifies individuals by birth and has historically determined access to social, economic, and cultural resources (Siddique, 2011). Caste continues to shape opportunities in education, employment, and social mobility. In contemporary governance, caste groups are broadly categorised as General, Other Backward Classes (OBCs), Scheduled Castes (SCs), and Scheduled Tribes (STs). SCs and STs, formally recognised under Article 341(1) of the Indian Constitution, have faced systemic discrimination and social exclusion (Kijima, 2006). To address these inequalities, the Indian government introduced affirmative action

policies (reservations), which have improved socioeconomic outcomes for marginalised communities (Siddique, 2011; Chin and Prakash, 2011).

Despite sustained legal and policy interventions, caste-based disparities persist, particularly in access to higher education and healthcare for Dalits and Adivasis (Thapa et al., 2021; Bhoi and Lakra, 2024). Increasingly, caste discrimination has also extended to digital spaces, where social media platforms host caste-based abuse that reflects and amplifies existing structural inequalities (and, 2025). Addressing these challenges requires computational approaches capable of detecting caste-based hate speech in culturally sensitive ways.

Automatic hate speech detection using machine learning (ML) and natural language processing (NLP) remains challenging due to linguistic diversity and cultural context (Abubakar et al., 2022). The challenge is particularly pronounced for Hindi, spoken by a large proportion of the Indian population but under-resourced for fine-grained hate speech analysis (Nautiyal, 2024). Existing Hindi hate speech datasets often rely on coarse binary labels, limiting their ability to capture target-specific abuse such as caste-based hate (Sharma et al., 2024). Furthermore, implicit references, coded language, and culturally embedded expressions make caste-related hate speech especially difficult to identify automatically.

Based on these gaps, we pose the following research questions:

- **RQ1:** How does a knowledge-driven, persona-based multi-agent framework affect the interpretability of caste-based hate speech classification?
- **RQ2:** Does the integration of a domain-specific Hindi caste hate lexicon improve the performance of multi-agent models in detecting caste-based hate speech?

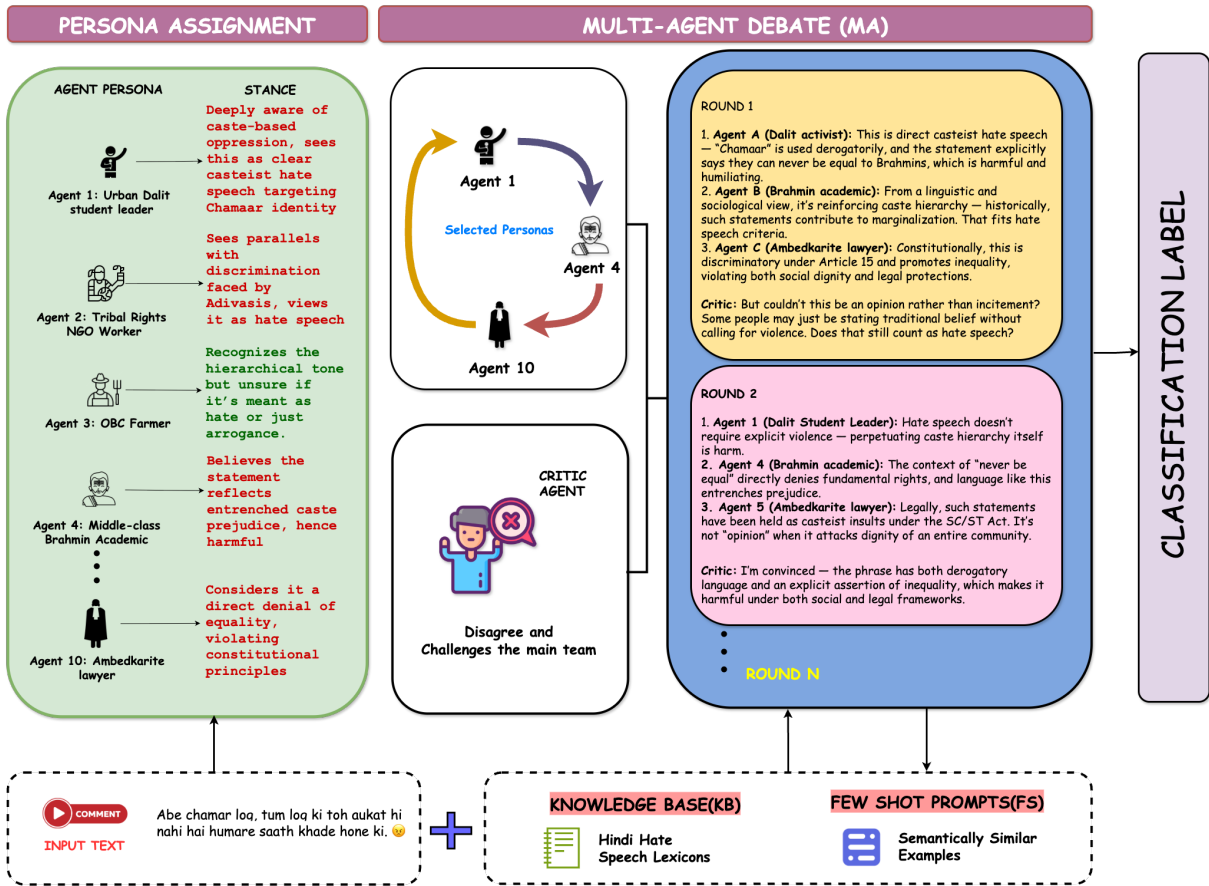


Figure 1: Overview of GYAAN-SAHIT for caste-based hate speech detection Framework

To address these questions, we propose GYAAN-SAHIT, a knowledge-driven, persona-based multi-agent framework for caste-based hate speech detection. It employs multiple agents with distinct socio-cultural perspectives and leverages structured, caste-specific lexicons to support culturally grounded reasoning.

The main contributions of this work are:

- **Hindi Caste Hate Lexicons:** Curated domain-specific lexicons capturing explicit slurs, implicit caste references, and culturally contextual expressions to support knowledge-grounded classification.
- **Knowledge-Driven Multi-Agent Framework:** A persona-based multi-agent architecture that improves reasoning in caste-based hate speech detection.
- **Comparative Analysis:** A systematic evaluation of transformer-based models, GPT-4o based prompting strategies, and multi-agent variants, demonstrating that knowledge-augmented persona-driven debate consistently

outperforms the baseline models and fine-tuned approaches.

2 Related Work

Most hate speech detection research has focused on high-resource languages, particularly English. Early approaches relied on rule-based and keyword matching methods, later superseded by fine-tuned transformer models such as BERT and RoBERTa (Davidson et al., 2017; Liu et al., 2019). However, models that perform well on curated evaluation datasets often generalise poorly to the diverse and informal environment of social media, where hate speech harms minority groups, fuels identity-based discrimination, and contributes to information disorder.

For low-resource languages, research has focused on domain adaptation and social media pre-training. The NAIJAHATE dataset for Nigerian Twitter demonstrates that standard pretrained models lack cultural specificity and require language-specific adaptation (Tonneau et al., 2024). The VIHATET5 model addresses Vietnamese data scarcity through multitask learning with noisy social me-

dia pretraining (Nguyen, 2024). Contrastive and adversarial learning have also been explored for other low-resource languages (Hashmi et al., 2025; Almahdi et al., 2025). Despite progress, significant challenges remain in handling socio-cultural diversity across platforms.

Most work on Indian language datasets targets general hate speech or abusive language rather than caste-based discrimination specifically (Mandl et al., 2019). A Tamil dataset focusing on caste and migration-based hate speech was recently released, and shared tasks such as LT-EDI 2024 addressed related problems (Chakravarthi et al., 2025; Rajiakodi et al., 2024). Resources and models for caste-based hate speech in Hindi remain scarce, and existing work is largely limited to coarse-grained, non-target-aware classification. Caste discourse often relies on implicit, coded, and indirect expressions that require deep understanding of social hierarchies. Large language models (LLMs) have been explored for hate speech detection in zero-shot and few-shot settings. While LLMs generalise well, single-agent prompting often yields unstable results, particularly in sensitive moderation tasks. Multi-agent and persona-driven frameworks address these limitations by improving robustness, control, and explainability. For example, the PREDICT framework engages multiple agents aligned with different annotation guidelines to produce more robust and explainable hate speech classifications (Park et al., 2024). Role-based debate among models has also been shown to improve reasoning in text generation tasks (Hu et al., 2024). Knowledge-driven multi-agent systems such as MAAD further demonstrate the value of structured domain knowledge for reliable and explainable AI (Zhang et al., 2025).

Despite these advances, no prior work has specifically addressed caste-based harmful discourse in Hindi using a culturally grounded, retrieval-augmented, multi-agent LLM framework. Existing approaches either fine-tune transformer models without structured interpretability or apply single-agent LLM prompting without deliberative reasoning or domain-specific lexical grounding. Our work addresses this gap by combining persona-driven multi-agent debate with curated caste-specific lexicons and retrieval-based contextual grounding to improve both performance and interpretability.

3 Methodology

Given an input comment t in Hindi, GYAANSAHIT outputs a final classification label \hat{L} . The framework comprises four stages: (1) persona assignment and selection, (2) multi-agent debate, (3) critic evaluation, and (4) final classification. All modules are implemented via LLM prompting, requiring no additional fine-tuning. An overview of the framework is provided in Figure 1, and Algorithm 1 presents the step-by-step procedure. Prompt configurations are detailed in Figures 5a–7 and sample outputs in Hindi and English are provided in Figures 8 and 9 in the Appendix.

3.1 Persona Assignment

Interpretations of sensitive language vary with an individual’s background and beliefs. To represent this diversity, we construct a pool of 6–10 personas, each representing a distinct socio-cultural perspective on caste-related discourse. This range captures a broad spectrum of viewpoints while avoiding redundancy. Using contextual cues from t and guided by socio-cultural diversity principles, we generate a set of caste perspective profiles:

$$C = \{c_1, c_2, \dots, c_n\} \quad (1)$$

Each perspective c_i is prompted to classify t with a label and a stance:

$$\begin{aligned} l_i &\in \{\text{Hatespeech}, \text{Not Hatespeech}\}, \\ s_i &= \text{reasoning for assigning } l_i \end{aligned} \quad (2)$$

Outputs are stored in a dictionary-style representation:

$$\begin{aligned} A = \{ \{ \text{agent_id}_1 : c_1, \text{label} : l_1, \text{stance} : s_1 \}, \\ \vdots \\ \{ \text{agent_id}_n : c_n, \text{label} : l_n, \text{stance} : s_n \} \} \end{aligned} \quad (3)$$

The personas are not intended to represent caste identities. Instead, they model diverse interpretive perspectives that may emerge in online discourse. Persona descriptions were designed to encourage diversity rather than simulate real individuals or communities.

Persona Pool Creation. Using LLM prompting, we generate perspectives that include a brief description, a classification label, and a reasoning stance, ensuring diversity across communities, ideologies, and viewpoints.

Algorithm 1 GYAAN-SAHIT: Knowledge-Based Multi-Agent Persona Framework for Caste-Based Hate Speech Classification

Require: Text t

Ensure: Final classification label $\hat{L} \in \{\text{Hatespeech, Not Hatespeech}\}$

- 1: Initialise agent pool $A \leftarrow \{\}$, selection meta-data $\text{Sel} \leftarrow \{\}$, debate state $D \leftarrow \{\}$
- 2: Generate caste perspective list $C = \{c_1, c_2, \dots, c_n\}$ from t
- 3: **for** each $c_i \in C$ **do**
- 4: Create persona for c_i ; prompt it to classify t , returning $(l_i^{(0)}, s_i^{(0)})$
- 5: Store: $A[\text{agent_id}_i] \leftarrow \{c_i, l_i^{(0)}, s_i^{(0)}\}$
- 6: **end for**
- 7: Select debate team $T = \{\text{agent_id}_1, \text{agent_id}_2, \text{agent_id}_3\}$
- 8: **for** each $\text{agent_id}_r \in T$ **do**
- 9: Record: $\text{Sel}[\text{agent_id}_r] \leftarrow \{l_r^{(0)}, \text{selection_criteria}\}$
- 10: Initialise: $D[\text{agent_id}_r] \leftarrow \{l_r^{(0)}, s_r^{(0)}\}$
- 11: **end for**
- 12: $k \leftarrow 0$
- 13: **repeat**
- 14: **for** each $\text{agent_id}_r \in T$ **do**
- 15: Agent presents current position $(l_r^{(k)}, s_r^{(k)})$
- 16: **end for**
- 17: Critic evaluates arguments: $F \leftarrow \mathcal{C}(\{(l_r^{(k)}, s_r^{(k)})\}_{\text{agent_id}_r \in T})$
- 18: **for** each $\text{agent_id}_r \in T$ **do**
- 19: Update: $(l_r^{(k+1)}, s_r^{(k+1)}) \leftarrow \text{UPDATE}(l_r^{(k)}, s_r^{(k)}, F)$
- 20: Write: $D[\text{agent_id}_r] \leftarrow \{l_r^{(k+1)}, s_r^{(k+1)}\}$
- 21: **end for**
- 22: $k \leftarrow k + 1$
- 23: **until** C is satisfied **or** $\text{AGREE}(\{(l_r^{(k)})\}_{\text{agent_id}_r \in T}) = \text{True}$
- 24: $\hat{L} \leftarrow \text{MAJORITYVOTE}(\{l_r^{(k)}\}_{\text{agent_id}_r \in T})$
- 25: **return** \hat{L}

Persona Selection. A debate team of three agents is selected from the pool:

$$T = \{a_j, a_k, a_l\} \subset A, \quad |T| = 3 \quad (4)$$

Three agents are sufficient to capture multiple, potentially conflicting perspectives; increasing this number offers diminishing returns and risks redun-

dancy. Each selected agent is assigned a fixed persona throughout the interaction to encourage consistent and balanced reasoning.

3.2 Multi-Agent Debate

Knowledge Integration via Retrieval-Augmented Guidance:

The framework integrates two complementary knowledge sources during debate.

Hindi Hate Speech Lexicons. We curated a domain-specific lexicon $\mathcal{L} = \{\ell_1, \ell_2, \dots, \ell_m\}$ of caste-related hate speech terms and implicit derogatory expressions in Hindi. An initial seed set was collected from online hate speech repositories including Hatebase¹, Homegrown², and the SC/ST Cell IIT Bombay³. This seed set was expanded using an LLM to generate additional terms and lexical variants. The finalised lexicon \mathcal{L} is provided to all agents as background knowledge, enhancing their ability to identify subtle and context-dependent hate speech during debate.

Retrieval-Augmented Guidance. To complement the lexicon, we employ a retrieval mechanism inspired by Retrieval-Augmented Generation (RAG), adapted for classification-oriented debate. All annotated training comments are embedded using a sentence embedding model and indexed in a FAISS vector database (Johnson et al., 2019). For a target comment x , the system retrieves the top- k most semantically similar examples:

$$\mathcal{N}_k(x) = \text{Top-}k(\text{sim}(\phi(x), \phi(x_i))) \quad (5)$$

along with their labels y_i . These $k=5$ retrieved pairs serve as contextual anchors, helping agents ground their reasoning in relevant precedents while reducing bias.

Debate Procedure:

The three selected agents $\mathcal{A} = \{a_1, a_2, a_3\}$ and a critic agent \mathcal{C} engage in structured, multi-turn debate to determine whether t constitutes caste-based hate speech. Each agent a_i interprets t using lexicon \mathcal{L} and the K retrieved examples $\mathcal{E} = \{(e_1, y_1), \dots, (e_K, y_K)\}$. At each debate round k , agent a_i formulates an argument comprising its classification label $l_i^{(k)}$ and supporting stance $s_i^{(k)}$, while challenging the reasoning of the other agents.

¹Hatebase.org

²Homegrown Voices: Casteist Slurs

³SC/ST Cell Resources

The critic \mathcal{C} evaluates the set of arguments:

$$\{(l_1^{(k)}, s_1^{(k)}), (l_2^{(k)}, s_2^{(k)}), (l_3^{(k)}, s_3^{(k)})\} \quad (6)$$

probing for logical weaknesses and inconsistencies. The debate proceeds for T rounds or until consensus is reached:

$$\text{Agree}\left(\{l_i^{(k)}\}_{i=1}^3\right) = \text{True} \quad (7)$$

The full procedure is outlined in Algorithm 1.

3.3 Critic Evaluation

Following the debate, the critic agent synthesises the discussion into a final classification decision, weighing argument coherence, the relevance of lexicon and retrieved evidence, and the overall balance of perspectives. The critic outputs:

$$\hat{L} \in \{\text{Hatespeech}, \text{Not Hatespeech}\} \quad (8)$$

together with a synthesised rationale integrating the strongest and most consistent arguments from the debate.

4 Experimental Setup

4.1 Dataset and Split

We use the dataset introduced by (Gupta et al., 2025), which underwent cleaning (removal of missing values, basic text normalisation, and outlier elimination) and numerical label encoding. Stratified sampling was used to split the data 80:20 into training and test sets, preserving the original class distribution.

We evaluate models under three settings. In the Seed setting, models use pre-trained weights fine-tuned on a small seed subset only, without access to the full training data. In the Train setting, models are fine-tuned on the 80% training split and evaluated on a stratified subset of 143 comments (73 hate, 70 non-hate). In the Full setting, models are fine-tuned on the complete dataset and evaluated on the full test set of 1,473 comments (712 hate, 761 non-hate). The Seed and Train settings assess performance under low-resource conditions, while the Full setting reflects performance at scale.

Transformer baselines and GPT-based multi-agent systems were evaluated under partially different experimental settings due to computational constraints associated with multi-agent inference. Therefore, comparisons across Tables 1 and 2 should be interpreted within-setting rather than as direct model-to-model superiority claims.

4.2 Baseline Models

Transformer Baselines. We evaluate four transformer-based models IndicBERT, MuRIL, RoBERTa-base, and XLM-RoBERTa-base under a unified architecture and consistent training protocol. Each model uses its pre-trained encoder with a dropout layer and a fully connected classification head applied to the [CLS] token. Fine-tuning was conducted for 3 epochs using the AdamW optimiser (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5} and batch size 16. Models were evaluated in both seed-data and full-training configurations.

GPT-4o Baselines. We evaluate GPT-4o under four prompting strategies without multi-agent mechanisms: zero-shot (ZS), where only the task description is provided; few-shot (FS), where a small set of labelled examples is included; seed data (SE), where human-annotated examples guide reasoning; and training data (TD), where curated examples enable stepwise classification decomposition.

4.3 Proposed Multi-Agent Variants

We implement four variants of GYAAN-SAHIT ZS, FS, SE, and TD mirroring the GPT-4o prompting conditions to enable direct comparison. In all variants, persona-driven agents are assigned distinct socio-cultural perspectives and engage in multi-turn debate. A critic agent assesses the evolving arguments for logical consistency and cultural appropriateness before producing the final label.

The full GYAAN-SAHIT-KB-FS model additionally incorporates retrieval-augmented guidance (top- k FAISS-based semantic retrieval) and the curated Hindi caste hate lexicon. Sentence embeddings are generated using paraphrase-multilingual-MiniLM-L12-v2 from SentenceTransformers; the annotated training set is embedded, indexed in FAISS, and linked to labels via a mapping file. All multi-agent reasoning is powered by GPT-4o. Retrieved examples were drawn exclusively from the training partition and not from evaluation samples. No test instances were indexed in the FAISS retrieval database.

5 Results and Discussion

5.1 Transformer Baseline Results

We evaluated four transformer models within a unified classification architecture. Across all three

Table 1: Performance of transformer-based models across training settings. **Seed**: pre-trained only, no task fine-tuning. **Train**: fine-tuned on the 80% split, evaluated on 143 comments (73 hate, 70 non-hate). **Full**: fine-tuned on the complete dataset, evaluated on 1,473 comments (712 hate, 761 non-hate). Bold values indicate the best scores per metric.

Model	Setting	Accuracy	Hatespeech			Not Hatespeech			Macro Avg			Weighted Avg		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
IndicBERT	Seed	0.49	0.00	0.00	0.00	0.49	1.00	0.66	0.24	0.50	0.33	0.24	0.49	0.32
	Train	0.78	1.00	0.56	0.72	0.69	1.00	0.81	0.84	0.78	0.77	0.85	0.78	0.77
	Full	0.79	0.93	0.61	0.73	0.72	0.96	0.82	0.82	0.78	0.78	0.82	0.79	0.78
RoBERTa	Seed	0.49	0.00	0.00	0.00	0.49	1.00	0.66	0.24	0.50	0.33	0.24	0.49	0.32
	Train	0.80	0.92	0.67	0.78	0.73	0.94	0.82	0.83	0.81	0.80	0.83	0.80	0.80
	Full	0.79	0.88	0.66	0.75	0.74	0.92	0.82	0.81	0.79	0.79	0.81	0.79	0.79
XLM-RoBERTa	Seed	0.49	0.00	0.00	0.00	0.49	1.00	0.66	0.24	0.50	0.33	0.24	0.49	0.32
	Train	0.81	0.91	0.70	0.79	0.75	0.93	0.83	0.83	0.81	0.81	0.83	0.81	0.81
	Full	0.81	0.83	0.77	0.80	0.80	0.86	0.83	0.82	0.81	0.81	0.81	0.81	0.81
MuRIL	Seed	0.49	0.00	0.00	0.00	0.49	1.00	0.66	0.24	0.50	0.33	0.24	0.49	0.32
	Train	0.87	0.90	0.84	0.87	0.84	0.90	0.87	0.87	0.87	0.87	0.87	0.87	0.87
	Full	0.83	0.85	0.79	0.82	0.81	0.87	0.84	0.83	0.83	0.83	0.83	0.83	0.83

settings (Table 1), Seed models consistently defaulted to the non-hate class, achieving zero recall for hate speech and confirming the necessity of task-specific fine-tuning regardless of model architecture.

In the Train setting (evaluated on 143 comments), MuRIL (Khanuja et al., 2021) achieved the strongest performance (0.87 accuracy, macro F1 0.87), benefiting from its pretraining on 17 Indian languages including code-mixed and transliterated text. XLM-RoBERTa-base (Conneau et al., 2020) followed with 0.81 accuracy and consistent F1 (0.81) across both classes. RoBERTa-base (Liu et al., 2019) reached 0.80 accuracy, while IndicBERT (Kakwani et al., 2020) achieved 0.78, with high precision (1.00) but low recall (0.56) for hate speech.

In the Full setting (evaluated on 1,473 comments), the ranking remained broadly consistent. MuRIL again led with 0.83 accuracy and balanced F1 (hate: 0.82, non-hate: 0.84). XLM-RoBERTa matched its Train performance at 0.81 accuracy. RoBERTa and IndicBERT both reached 0.79 accuracy, with IndicBERT retaining its pattern of high precision (0.93) but low hate recall (0.61). The convergence of Train and Full results for most models suggests that the 80% training split captures sufficient distributional diversity for effective fine-tuning.

t-SNE visualisations (Figures 2 and 3) corroborate these findings: Seed models show poor class separation, while fine-tuned models form clearly distinct hate and non-hate clusters.

5.2 GPT-Based vs. Proposed Multi-Agent Models

Among GPT-4o-only baselines, GPT-TD achieved the highest accuracy (0.79) with balanced precision and recall. GPT-FS exhibited perfect precision (1.00) for hate speech but extremely low recall (0.08), reflecting over-reliance on a narrow example set. GPT-ZS and GPT-SE underperformed with accuracies of 0.45 and 0.47 respectively.

Introducing multi-agent debate improved performance in most configurations. GYAAN-SAHIT-ZS raised accuracy from 0.45 to 0.79, substantially increasing hate speech recall. GYAAN-SAHIT-FS, SE, and TD demonstrated more balanced precision-recall trade-offs than their single-agent counterparts, highlighting the benefit of persona diversity and collaborative reasoning.

GYAAN-SAHIT-KB-FS achieved the best results across all experiments as shown in Figure 4 and Table 2: 0.8881 accuracy, macro/weighted F1 of 0.89, high hate speech precision (0.98) with recall (0.79), and non-hate performance (precision 0.99, recall 0.82).

6 Ablation Study

To assess the contribution of individual components in GYAAN-SAHIT, we conducted an ablation study removing the critic agent and fixing the debate to 1 to 3 rounds, while retaining retrieval-augmented guidance ($k=10$ retrieved examples via FAISS-indexed paraphrase-multilingual-MiniLM-L12-v2 embeddings) and few-shot prompting. Table 3 shows the results of ablation study.

The full model (with critic) achieved the highest

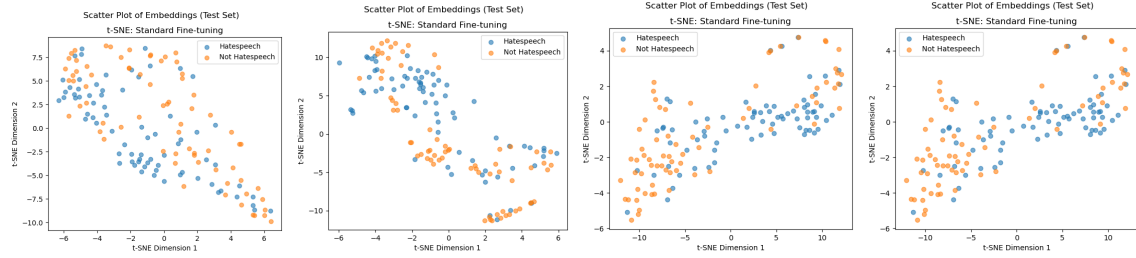


Figure 2: t-SNE visualisations of seed-data models: XLM-RoBERTa, RoBERTa, MuRIL, and IndicBERT.

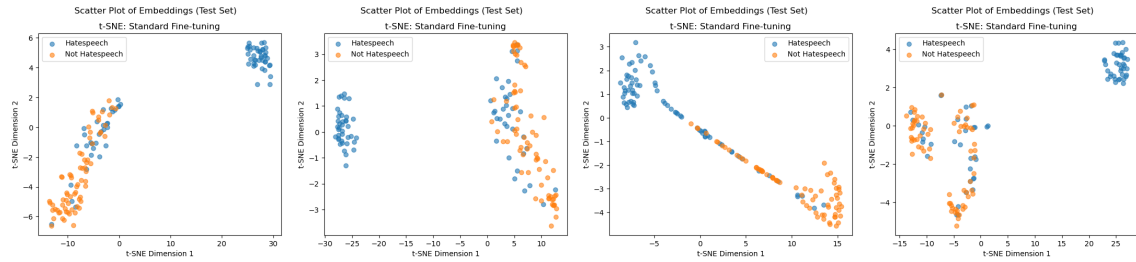


Figure 3: t-SNE visualisations of train-data models: XLM-RoBERTa, RoBERTa, MuRIL, and IndicBERT.

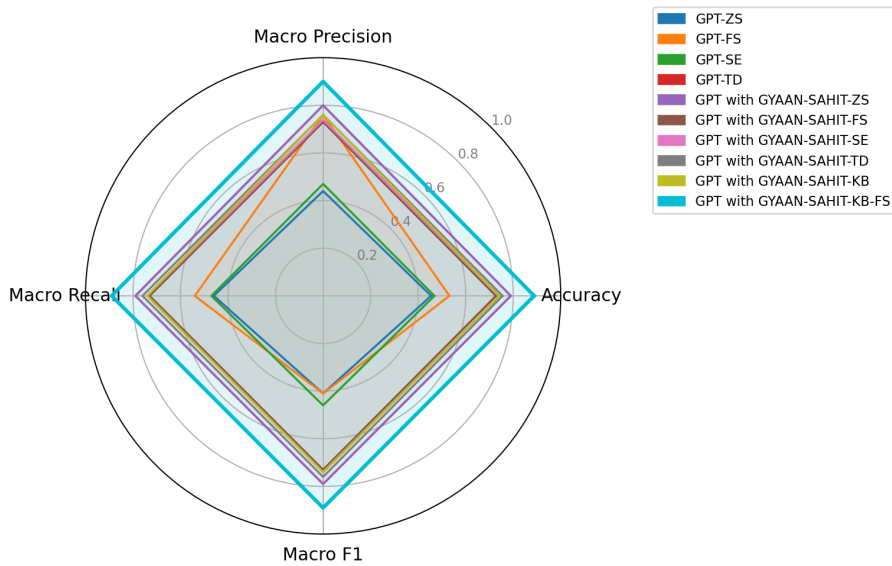


Figure 4: Radar chart comparing GPT-based and proposed multi-agent models on accuracy, macro precision, macro recall, and macro F1.

Table 2: Performance of GPT-4o baselines and corresponding GYAAN-SAHIT variants on the test set.

Model	Accuracy	Hate			Non-Hate			Macro Avg			Weighted Avg		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-ZS	0.4545	0.41	0.16	0.24	0.46	0.76	0.58	0.44	0.46	0.41	0.44	0.45	0.40
GPT with GYAAN-SAHIT-ZS	0.7902	0.84	0.73	0.78	0.75	0.86	0.80	0.80	0.79	0.79	0.80	0.79	0.79
GPT-FS	0.5315	1.00	0.08	0.15	0.51	1.00	0.68	0.76	0.54	0.41	0.76	0.53	0.41
GPT with GYAAN-SAHIT-FS	0.7273	0.74	0.71	0.73	0.71	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73
GPT-SE	0.4685	0.47	0.34	0.40	0.47	0.60	0.53	0.47	0.47	0.46	0.47	0.47	0.46
GPT with GYAAN-SAHIT-SE	0.7413	0.74	0.75	0.75	0.74	0.73	0.73	0.74	0.74	0.74	0.74	0.74	0.74
GPT-TD	0.7902	0.88	0.68	0.77	0.73	0.90	0.81	0.80	0.79	0.79	0.81	0.79	0.79
GPT with GYAAN-SAHIT-TD	0.7552	0.76	0.75	0.76	0.75	0.76	0.75	0.76	0.76	0.76	0.76	0.76	0.76
GPT - KB	0.7413	0.70	0.86	0.77	0.81	0.61	0.70	0.76	0.74	0.74	0.75	0.74	0.74
GPT with GYAAN-SAHIT-KB-FS	0.8881	0.98	0.79	0.88	0.82	0.99	0.90	0.90	0.89	0.89	0.90	0.89	0.89

Table 3: Ablation results for GYAAN-SAHIT: effect of removing the critic agent and varying debate rounds.

Model	Accuracy	Hate			Non-Hate			Macro Avg			Weighted Avg		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
GYAAN-SAHIT-KB-FS with critic (n rounds)	0.7667	0.92	0.67	0.77	0.65	0.92	0.76	0.79	0.79	0.77	0.81	0.77	0.77
w/o critic — 1 round	0.7333	0.86	0.67	0.75	0.62	0.83	0.71	0.74	0.75	0.73	0.76	0.73	0.74
w/o critic — 2 rounds	0.7333	0.86	0.67	0.75	0.62	0.83	0.71	0.74	0.75	0.73	0.76	0.73	0.74
w/o critic — 3 rounds	0.7333	0.86	0.67	0.75	0.62	0.83	0.71	0.74	0.75	0.73	0.76	0.73	0.74

accuracy (0.7667) and outperformed all ablation variants across precision, recall, and F1. All three ablation configurations regardless of the number of rounds produced identical results (0.7333 accuracy), indicating that without the critic, iterative debate alone does not improve classification. These findings confirm that both the critic’s evaluative role and extended debate contribute meaningfully to performance. The consistent results across 1–3 rounds also suggest that retrieval augmented guidance remains the dominant driver of quality; the critic provides additional refinement, particularly for ambiguous cases.

Conclusion

We introduced GYAAN-SAHIT-KB-FS, a knowledge-augmented multi-agent framework for caste-based hate speech detection in Hindi. By combining few-shot prompting, FAISS-based retrieval-augmented guidance, a curated Hindi caste hate lexicon, and structured persona-driven debate with critic evaluation, the framework demonstrates competitive performance relative to transformer and GPT-based inference models within the evaluated setting. On the test set, GYAAN-SAHIT-KB-FS achieves 0.8881 accuracy and balanced macro F1 of 0.89. The future work is extending GYAAN-SAHIT to additional low-resource Indian languages and evaluating multilingual lexicon adaptation strategies.

Ethical Statement

The proposed framework addresses socially sensitive content involving caste-based discrimination and hate speech. Although persona-driven reasoning can improve interpretability, it may also risk reinforcing socio-cultural stereotypes if poorly designed. To mitigate this risk, personas were constructed around interpretive roles rather than fixed caste identities. The framework is intended as a decision-support system and not as a fully autonomous moderation pipeline. Human oversight remains important, particularly for borderline cases

involving reclaimed language, sarcasm, or context-dependent expressions.

Limitations

Although GYAAN-SAHIT demonstrates promising performance for caste-based hate speech detection, the current study has several limitations. Due to computational constraints, we did not conduct extensive bootstrap confidence interval estimation or paired significance testing in the current version. Evaluation is also limited to Hindi caste-related discourse and may not generalise directly to other languages or sociocultural settings. In addition, the persona-based debate framework depends on GPT-4o prompting, which introduces additional computational cost and inference latency. The curated lexicon may also contain incomplete or context-sensitive terms that require continuous validation.

Acknowledgment

Bharathi Raja Chakravarthi and Shunmuga Priya Muthusamy Chinnan were funded by Research Ireland under grant number SFI/12/RC/2289_P2 (Insight_2).

References

- Ibrahim Abubakar, Lu Gram, Sarah Lasoye, E Tendam Achime, Laia Becares, Gurpreet Kaur Bola, Rageshri Dhairyawan, Gideon Lasco, Martin McKee, Yin Paradies, Nidhi S Sabharwal, Sujitha Selvarajah, Geordan Shannon, and Delan Devakumar. 2022. [Confronting the consequences of racism, xenophobia, and discrimination on health and health-care systems](#). *The Lancet*, 400(10368):2137–2146.
- Asseel Jabbar Almahdi, Ali Mohades, Mohammad Akbari, and Soroush Heidary. 2025. [Enhancing cross-lingual hate speech detection through contrastive and adversarial learning](#). *Engineering Applications of Artificial Intelligence*, 147:110296.
- Dhyan Singh and. 2025. [Dalits’ encounters with casteism on social media: a thematic analysis](#). *Information, Communication & Society*, 28(2):335–353.
- Dhaneswar Bhoi and Neelima Rashmi Lakra. 2024. [Creating cultural capital for dalits and adivasis: Policy](#)

- passage for their de-marginalisation in higher education. *Journal of Social Inclusion Studies*, 10(1):98–119.
- Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Sara Yogesh Thakare, and Sathiyaraj Thangasamy. 2025. [Detecting caste and migration hate speech in low-resource tamil language](#). *Language Resources and Evaluation*.
- Aimee Chin and Nishith Prakash. 2011. [The redistributive effects of political reservation for minorities: Evidence from india](#). *Journal of Development Economics*, 96(2):265–277.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.
- Sakshi Gupta, Shunmuga Priya Muthusamy Chinnan, Saranya Rajiakodi, Ratnavel Rajalakshmi, Rahul Ponnusamy, and Bharathi Raja Chakravarthi. 2025. [Caste-based hate speech detection in low-resource hindi language](#). In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, DHOW '25, page 55–64, New York, NY, USA. Association for Computing Machinery.
- Ehtesham Hashmi, Sule Yildirim Yayilgan, Muhammad Mudassar Yamin, Mohamed Abomhara, and Mohib Ullah. 2025. [Self-supervised hate speech detection in norwegian texts with lexical and semantic augmentations](#). *Expert Systems with Applications*, 264:125843.
- Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2024. [Debate-to-write: A persona-driven multi-agent framework for diverse argument generation](#). *arXiv preprint arXiv:2406.19643*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. [Billion-scale similarity search with gpus](#). *IEEE Transactions on Big Data*.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLPsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vishnu Subramanian, and Partha Pratim Talukdar. 2021. [Muril: Multilingual representations for indian languages](#). *ArXiv*, abs/2103.10730.
- Yoko Kijima. 2006. [Caste and tribe inequality: Evidence from india, 1983-1999](#). *Economic Development and Cultural Change*, 54(2):369–404.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations (ICLR)*.
- Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. [Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages](#). In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19*, page 14–17, New York, NY, USA. Association for Computing Machinery.
- Dr. Jayanti Prasad Nautiyal. 2024. [Hindi, the language of the largest population in the world—facts and figures \(research report 2023—revised for ethnologue world language ranking\)](#). *International Journal of All Subject Research*, 3(1):16–29.
- Luan Nguyen. 2024. [Vihatet5: Enhancing hate speech detection in vietnamese with a unified text-to-text transformer model](#). pages 5948–5961.
- Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. [PREDICT: Multi-agent-based debate simulation for generalized hate speech detection](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987, Miami, Florida, USA. Association for Computational Linguistics.
- Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. [Overview of shared task on caste and migration hate speech detection](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 145–151, St. Julian's, Malta. Association for Computational Linguistics.
- Deepawali Sharma, Vivek Kumar Singh, and Vedika Gupta. 2024. [Tabhate: A target-based hate speech detection dataset in hindi](#). *Social Network Analysis and Mining*, 14(1):190.

Zahra Siddique. 2011. [Evidence on caste based discrimination](#). *Labour Economics*, 18:S146–S159. Labour markets in developing countries.

Raksha Thapa, Edwin van Teijlingen, Pramod Raj Regmi, and Vanessa Heaslip. 2021. [Caste exclusion and health discrimination in south asia: A systematic review](#). *Asia Pacific Journal of Public Health*, 33(8):828–838. PMID: 34024157.

Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. [Naija-Hate: Evaluating hate speech detection on Nigerian Twitter using representative data](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9020–9040, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhang, Ruiyin Li, Peng Liang, Weisong Sun, and Yang Liu. 2025. [Knowledge-based multi-agent framework for automated software architecture design](#). *Preprint*, arXiv:2503.20536.

Appendix

Prompt 1: Persona Creation Prompt	
इनपुट Input	इनपुट टेक्स्ट Input text
ज़िम्मेदारी Responsibility	6 से 10 पर्सनालिटी एजेंट बनाएं Create a 6 to 10 Personas agents
उद्देश्य Objective	लक्ष्य 6-10 वाद-विवाद व्यक्तित्व एजेंटों का एक समूह तैयार करना है, जिनमें से प्रत्येक को दिए गए पाठ का एक विशिष्ट दृष्टिकोण से विश्लेषण करने का कार्य सौंपा गया है। प्रत्येक एजेंट को एक विशिष्ट दृष्टिकोण प्रस्तुत करना होगा, जो इस विचार का समर्थन या विरोध करेगा कि पाठ में जाति-आधारित घृणा है, और यह सुनिश्चित करना होगा कि इनपुट से संबंधित सभी दृष्टिकोणों का प्रतिनिधित्व किया जाए। The goal is to create a set of 6-10 debate persona agents, each tasked with analysing the given text from a distinct perspective. Every agent should embody a unique viewpoint, either supporting or opposing the idea that the text contains caste-based hate, ensuring that all perspectives relevant to the input are represented.
प्रतिबंध Constraints	1. प्रत्येक एजेंट के लिए, एक वाक्य में वर्णित एक सामान्य जाति-संबंधी व्यक्तित्व प्रदान करें। 2. इनपुट टेक्स्ट के मूल्यांकन के लिए एक विशिष्ट विश्लेषणात्मक दृष्टिकोण को परिभाषित करने वाला एक संगत दावा निर्दिष्ट करें। 3. सुनिश्चित करें कि प्रत्येक दावा और प्रतिक्रिया विशिष्ट हो और एक अद्वितीय दृष्टिकोण पर केंद्रित हो। 4. जाति-आधारित घृणास्पद भाषण का पता लगाने के लिए सभी व्यक्तित्वों और दावों को प्रासंगिक रखें। For each agent, provide a generic caste-related persona described in one sentence. Assign a corresponding claim that defines a specific analytical angle for evaluating the input text. Ensure each claim and response is distinct and focuses on a unique perspective. Keep all personas and claims relevant to caste-based hate speech detection.
आउटपुट स्वरूप Output Format	आउटपुट का प्रारूप JSON में होना चाहिए { "एजेंट_आईडी": "सारांश": "एजेंट का सारांश", "वर्गीकरण_लेबल": "एजेंट का वर्गीकरण_लेबल" } . Format of the output should be in Json { "agent_id": "summary": "the_summary_of_Agent", "classification_label": "the_classification_label_of_Agent" } .

(a) Prompt 1: Persona Creation. Generates diverse socio-cultural personas with distinct viewpoints on caste-related issues, forming the foundation of the multi-agent debate.

Prompt 2: Persona Pool Selection Prompt	
इनपुट Input	प्रॉम्प्ट 1 से वर्गीकरण लेबल के साथ पाठ इनपुट करें। Input text with classification label from Prompt 1
ज़िम्मेदारी Responsibility	3 पर्सोना एजेंट का चयन करें। Select the 3 Personas agents
उद्देश्य Objective	कार्य तीन एजेंटों की एक टीम बनाना है, जो एक साथ काम करें और सहयोगात्मक रूप से एक प्रेरक वर्गीकरण लेबल तैयार करें जो दिए गए पाठ का विश्लेषण करे। The task is to build a team of three agents, to work together and collaboratively formulate a persuasive classification_label that analyse the given text.
प्रतिबंध Constraints	<ul style="list-style-type: none"> प्रत्येक उम्मीदवार का व्यक्तित्व अद्वितीय होता है और वह विषय पर एक अलग दृष्टिकोण प्रस्तुत करता है। तीन ऐसे एजेंटों का चयन करें जो सामूहिक रूप से कार्य के लिए एक मजबूत और प्रभावी टीम बना सकें। संतुलित और निष्पक्ष चर्चा को बढ़ावा देने के लिए दृष्टिकोणों की विविधता सुनिश्चित करें। प्रत्येक चयनित एजेंट के लिए, चयन मानदंड प्रदान करें और स्पष्ट करें कि उन्हें क्यों चुना गया। Each candidate has a unique persona offering a different perspective on the topic. Select three agents that can collectively form a strong, effective team for the task. Ensure diversity of perspectives to promote a balanced and fair discussion. For each selected agent, provide selection criteria explaining why they were chosen.
आउटपुट स्वरूप Output Format	आउटपुट का प्रारूप JSON में होना चाहिए { "एजेंट_आईडी": "सारांश": "एजेंट का विवरण", "वर्गीकरण_लेबल": "एजेंट का वर्गीकरण_लेबल", "चयन_मानदंड": "चयन_मानदंड" } . Format of the output should be in Json { "agent_id": "summary": "the_description_of_Agent", "classification_label": "the_classification_label_of_Agent", "selection_criteria": "selection_criteria" } .

(b) Prompt 2: Persona Pool Selection. Selects a three-agent debate team from the generated pool, ensuring diversity of perspectives and complementary stances.

Figure 5: Prompts used in the persona generation and selection stages of the proposed framework.

Prompt 4: Classification Label Prompt	
इनपुट Input	प्रॉम्प्ट 3 से सिमुलेशन विवरण Simulation Details from Prompt 3
ज़िम्मेदारी Responsibility	वर्गीकरण लेबल प्रदान करें Provide the classification Label
उद्देश्य Objective	वर्गीकरण लेबल प्रदान करें कार्य वाद-विवाद सिमुलेशन विवरण के अनुसार वर्गीकरण लेबल प्रदान करना है। The task is to provide the classification label according to the debate simulation details.
आउटपुट स्वरूप Output Format	वर्गीकरण लेबल Classification label

Figure 7: Prompt 4: Final Classification Label Generation.

Prompt 3: Debate Simulation Prompt	
इनपुट Input	<p>प्रॉम्प्ट 2 से 3 एजेंटों का चयन किया गया। जानकोष (हिंदी जातिस्थक अपशब्दों का शब्दकोश) Selected 3 agent's from Prompt 2. Knowledge base(Hindi Caste slurs lexicons)</p> <p>कुछ शॉट (k=5) उदाहरण (FAISS और RAG का उपयोग करके अर्थगत रूप से समान)</p> <ol style="list-style-type: none"> "अबे चमार लोग, तुम्हारी औकात क्या है? तुम्हें तो बस झाड़ू लगाना चाहिए! 🗡️" → हेटस्पीच "देखो, धोबी को तो बस सफाई का काम करना चाहिए, क्या फालतू की बातें करते हैं! 🗡️" → हेटस्पीच. . . <p>Few Shot (k=5) examples (Semantically similar using FAISS with RAG)</p> <ol style="list-style-type: none"> "Abe chamar log, tumhari aukaat kya hai? Tumhe toh bas jhaadu lagana chahiye! 🗡️" → Hatespeech "Dekho, in dhobi ko toh bas safai ka kaam karna chahiye, kya faltu ki baatein karte hain. 🗡️" → Hatespeech . .
ज़िम्मेदारी Responsibility	<p>एक बहस प्रक्रिया का अनुकरण करें. Simulate a debate process.</p>
उद्देश्य Objective	<p>कार्य जाति आधारित विषय पर दिए गए पाठ का विश्लेषण करना, तथा वाद-विवाद चर्चा के आधार पर पाठ का विश्लेषण करने के लिए एक सुव्यवस्थित वर्गीकरण लेबल योजना तैयार करना है।</p> <p>The task is to analyse a given text on a caste based topic, and formulate a well-structured classification label plan to analyse the text based on the debate discussion.</p>
निर्देश Instructions	<ol style="list-style-type: none"> तीन सदस्यों की एक मुख्य टीम: एजेंट A, एजेंट B, और एजेंट C <ul style="list-style-type: none"> -रुख: दिए गए पाठ का विश्लेषण करें; -लक्ष्य: पाठ का विश्लेषण करने के लिए समग्र तार्किक प्रवाह की रूपरेखा तैयार करते हुए एक प्रेरक वर्गीकरण-लेबल योजना प्रस्तावित करने के लिए एक साथ चर्चा करें। -टीम के सदस्यों के विशिष्ट व्यक्तित्व और वर्गीकरण-लेबल: <ul style="list-style-type: none"> -एजेंट A: व्यक्तित्व-क; वर्गीकरण-लेबल; -एजेंट B: व्यक्तित्व-ख; वर्गीकरण-लेबल; -एजेंट C: व्यक्तित्व-ग; वर्गीकरण-लेबल; एक आलोचक <ul style="list-style-type: none"> -रुख: दिए गए पाठ का विश्लेषण करें; -लक्ष्य: आप मुख्य टीम से असहमत हैं। मुख्य टीम की चर्चा में कमज़ोरियों की पहचान करें और उन्हें चुनौती दें, और मुख्य टीम के साथ बहस करें। <ol style="list-style-type: none"> A Main Team of three members: Agent A, Agent B, and Agent C <ul style="list-style-type: none"> - Stance: Analyse the Given text; - Goal: Discuss together to propose a persuasive classification_label plan outlining the overall logical flow to analyse text. - Specific Personas and classification_label of the team members: <ul style="list-style-type: none"> - Agent A: persona_a; classification_label; - Agent B: persona_b; classification_label; - Agent C: persona_c; classification_label; A Critic <ul style="list-style-type: none"> - Stance: Analyse the Given text; - Goal: You Disagree with the Main Team. Identify and challenge weaknesses in the Main Team's discussion, and debate with the Main team.
प्रतिबंध Constraints	<ul style="list-style-type: none"> मुख्य टीम के वर्गीकरण लेबल पर सहमत होने और आलोचक को सहमत होने तक कई दौर तक चर्चा चलाएँ। तर्क-वितर्क को सुदृढ़ बनाएँ ताकि तर्क ठोस और सुसंगत हो। बोलने के किसी सख्त क्रम की आवश्यकता नहीं है, लेकिन विषयों का क्रम एक जैसा होना चाहिए। Conduct the discussion for multiple rounds until the Main Team agrees on the classification_label and the Critic is persuaded. Ensure rigorous reasoning so that the logic is persuasive and coherent. No strict speaking order is required, but the sequence of topics should maintain a clear and logical progression.
आउटपुट स्वरूप Output Format	<p>सिम्युलेशन विवरण Simulation details</p>

Figure 6: Prompt 3: Multi-Round Debate. Guides persona interaction, argument exchange, iterative stance refinement, and critic-based synthesis to produce a logically coherent and culturally informed final label.

Table 4: Model predictions and explanations for a Not Hatespeech example.

Model	Label	Explanation
Input (Code-mix): <i>Dhere dhere hindu dharam bhi half ho jayega</i>		
Translation: “Hindu religion will also be halved.”		
True Label: Not Hatespeech		
GPT-ZS	Not Hatespeech	Expresses concern about Hinduism’s decline but does not target any caste group.
GPT-FS	Not Hatespeech	A general religious opinion with no caste-specific hate content.
GPT-SE	Not Hatespeech	A general statement about religion; no caste is referenced or attacked.
GPT-TD	Not Hatespeech	Vague concern about a religion’s future with no caste reference or target.
GYAAN-SAHIT-ZS	Not Hatespeech	Lacks direct incitement, slurs, or hostility; interpretable as pessimistic commentary.
GYAAN-SAHIT-FS	Not Hatespeech	References Hinduism generally without targeting or insulting any caste group.
GYAAN-SAHIT-SE	Not Hatespeech	A pessimistic religious statement without derogatory caste or religious terms.
GYAAN-SAHIT-TD	Not Hatespeech	A prediction about social change with no derogatory terms or calls for harm.
GYAAN-SAHIT-KB	Hatespeech	Hostility toward a religious community predicting its decline aligns with hate patterns.
GYAAN-SAHIT-KB-FS	Not Hatespeech	Predicts religious decline without derogatory language or calls for harm.

Table 5: Model predictions and explanations for a Hatespeech example.

Model	Label	Explanation
Input (Code-mix): <i>Upper caste waale toh bas baniya logon ko hi support karte hain. Hum dalit toh bas vote bank hain.</i>		
Translation: “The upper caste people only support the Baniyas. We Dalits are just a vote bank.”		
True Label: Hatespeech		
GPT-ZS	Not Hatespeech	Expresses frustration about caste politics but does not incite violence.
GPT-FS	Not Hatespeech	Voices political disenchantment without slurs or incitement.
GPT-SE	Hatespeech	Sweeping generalisation about upper castes reinforces caste divisions.
GPT-TD	Hatespeech	Generalises caste groups as biased and exploitative.
GYAAN-SAHIT-ZS	Not Hatespeech	Critiques political dynamics without inciting hatred.
GYAAN-SAHIT-FS	Hatespeech	Contains harmful generalisations reinforcing divisive narratives.
GYAAN-SAHIT-SE	Not Hatespeech	Critiques caste favouritism without expressing hate.
GYAAN-SAHIT-TD	Hatespeech	Reflects caste-based hostility through derogatory generalisations.
GYAAN-SAHIT-KB	Hatespeech	Stereotypes upper castes and Baniyas with negative intent.
GYAAN-SAHIT-KB-FS	Hatespeech	Targets caste identity through sweeping negative generalisation.

<p>इनपुट टेक्स्ट दलित के नाम पर गुण्डों की पहचान कर उन्हें जहनम भेजो </p>
<p>व्यक्तित्व एजेंट</p> <ol style="list-style-type: none"> एजेंट सविता (दलित महिला अधिकार कार्यकर्ता): लेबल: घृणास्पद भाषण रुख: यह बयान जाति से पहचाने जाने वाले लोगों के खिलाफ हिंसा का आह्वान करता है, जो सीधे तौर पर घृणास्पद भाषण है। एजेंट राघव (उच्च जाति के शिक्षाविद): लेबल: घृणास्पद भाषण रुख: यह "दलित" को नकारात्मक रूप में दर्शाता है और उन्हें "नरक" में भेजने की वकालत करता है, जिससे शत्रुता को बढ़ावा मिलता है। एजेंट प्रकाश (ओबीसी किसान नेता): लेबल: घृणास्पद भाषण रुख: यह वाक्यांश एक जाति समूह के लिए खुले तौर पर धमकी भरा और हिंसक है। एजेंट मीना (दलित नारीवादी विद्वान): लेबल: घृणास्पद भाषण रुख: भले ही यह "गुंडों" के लिए लक्षित हो, लेकिन हिंसा के आह्वान में जाति को पहचान के रूप में इस्तेमाल करना भेदभावपूर्ण है। एजेंट अरविंद (ब्राह्मण पत्रकार): लेबल: घृणास्पद भाषण रुख: यह बयान जाति के लेबल को हिंसक बयानबाजी के साथ मिलाता है, जो सार्वजनिक विमर्श में खतरनाक है। एजेंट शालिनी (अंबेडकरवादी छात्रा) लेबल: घृणास्पद भाषण रुख: यह जाति की पहचान को अपराध से गलत तरीके से जोड़ता है, जो एक पारंपरिक घृणास्पद रूपक है। एजेंट फरहान (अल्पसंख्यक अधिकारों के लिए काम करने वाले पर्यवेक्षक): लेबल: घृणास्पद भाषण रुख: यह निर्माण "जहनुम भेजो" के माध्यम से घृणा और अमानवीयकरण को बढ़ाता है। एजेंट निर्मला (अनुसूचित जाति की सामाजिक कार्यकर्ता): लेबल: घृणास्पद भाषण रुख: यह जाति से जुड़ी पहचान के आधार पर व्यक्तियों को नुकसान पहुंचाने का एक स्पष्ट आह्वान है।
<p>3 चयनित व्यक्तित्व एजेंट</p> <ol style="list-style-type: none"> एजेंट सविता - दलित अधिकारों के दृष्टिकोण से परिप्रेक्ष्य प्रस्तुत करती हैं, जाति-हिंसा संबंधी बयानबाजी के नुकसान को पहचानती हैं। एजेंट राघव - अकादमिक विश्लेषणात्मक दृष्टिकोण, खतरनाक भाषाई संरचना पर जोर देती हैं। एजेंट शालिनी - इस बात पर गहरी कार्यकर्ता अंतर्दृष्टि कि कैसे ऐसे बयान प्रणालीगत भेदभाव को बढ़ावा देते हैं।
<p>बनावटी बहस</p> <p>पहला दौर</p> <ol style="list-style-type: none"> सविता: यह सीधे तौर पर "दलित" को एक नकारात्मक पहचानकर्ता के रूप में इस्तेमाल करता है और उन्हें "नरक" में भेजने की वकालत करता है, जो स्पष्ट रूप से हिंसक घृणास्पद भाषण है। राघव: भले ही "गुंडे" शब्द का उद्देश्य लक्ष्य को सीमित करना हो, फिर भी इसे जाति के नाम के साथ जोड़ने से उस जाति के प्रति हिंसा का सामान्यीकरण हो जाता है। शालिनी: ऐतिहासिक रूप से, ऐसी भाषा का इस्तेमाल हाशिए पर पड़े समुदायों पर हमलों को सही ठहराने के लिए किया जाता रहा है। यह घृणा बढ़ाने का एक उत्कृष्ट उदाहरण है। <p>आलोचक: क्या वक्ता "दलित के नाम पर गुंडे" (यानी, छद्मवेशी) कह रहा है, न कि वास्तविक दलित?</p> <p>दूसरा दौर</p> <ol style="list-style-type: none"> सविता: फिर भी, यह कथन पहचान को अपराध से अलग नहीं कर पाता; यह अभी भी जनता की धारणा में पूर्वाग्रह को समाहित करता है। राघव: भाषाई अस्पष्टता नुकसान को कम नहीं करती—अगर लोग इसे आम तौर पर दलितों के खिलाफ समझते हैं, तो यह खतरनाक है। शालिनी: और यहाँ, "जहनुम भेजो" स्पष्ट रूप से नुकसान पहुंचाने का आह्वान है, जो इसे इरादे की बारीकियों के बावजूद अस्वीकार्य बनाता है। <p>आलोचक: तो आपका निर्णय शब्दों और दर्शकों की संभावित धारणा, दोनों पर आधारित है?</p> <p>तीसरा दौर</p> <ol style="list-style-type: none"> सविता: हाँ, क्योंकि सामाजिक संदर्भों में भाषा वास्तविक दुनिया के कार्यों को आकार देती है। राघव: विशेष रूप से जाति-संवेदनशील वातावरण में, हिंसक वाक्यांश स्वाभाविक रूप से उकसाने वाले होते हैं। शालिनी: इसलिए, यह स्पष्ट रूप से घृणास्पद भाषण माना जाता है। <p>अंतिम वर्गीकरण लेबल - अभद्र भाषा</p>

Figure 8: Example output from GYAAN-SAHIT for caste-based hate speech classification, shown in the original Hindi.

<p>Input text Identify the goons in the name of Dalits and send them to hell.</p>
<p>Persona Agents</p> <ol style="list-style-type: none"> Agent Savita (Dalit woman rights activist): Label: Hatespeech Stance: The statement calls for violence against people identified by caste, which is direct hate speech. Agent Raghav(Upper-caste academic): Label: Hatespeech Stance: It singles out "Dalit" in a negative frame and advocates sending them to "hell," promoting hostility. Agent Prakash (OBC farmer leader): Label: Hatespeech Stance: The phrase is openly threatening and violent toward a caste group. Agent Meena (Dalit feminist scholar): Label: Hatespeech Stance: Even if aimed at "goons," using caste as an identifier in a call for violence is discriminatory. Agent Arvind (Brahmin journalist): Label: Hatespeech Stance: This statement blends a caste label with violent rhetoric, dangerous in public discourse. Agent Shalini (Ambedkarite student) Label: Hatespeech Stance: It wrongly associates caste identity with criminality, a classic hate trope. Agent Farhan (Minority rights observer): Label: Hatespeech Stance: The construction incites hatred and dehumanization through "jahannum bhejo." Agent Nirmala (SC social worker): Label: Hatespeech Stance: It's an unambiguous call to harm individuals based on caste-linked identification.
<p>3 Selected Persona Agents</p> <ol style="list-style-type: none"> Agent Savita - Brings perspective from a Dalit rights standpoint, recognizing the harm of caste-violence rhetoric. Agent Raghav - Academic analytical approach, emphasizing the dangerous linguistic structure. Agent Shalini - Strong activist insight into how such statements fuel systemic discrimination.
<p>Simulated debate</p> <p>Round 1</p> <ol style="list-style-type: none"> Savita: This directly uses "Dalit" as a negative identifier and advocates sending them to "hell," which is clearly violent hate speech. Raghav: Even if the term "goons" was meant to limit the target, coupling it with a caste name still generalizes violence toward that caste. Shalini: Historically, such language has been used to justify attacks on marginalized communities. This is classic hate incitement. <p>Critic: Could the speaker be saying "goons in the name of Dalit" (i.e., impersonators) and not actual Dalits?</p> <p>Round 2</p> <ol style="list-style-type: none"> Savita: Even then, the statement fails to separate the identity from criminality; it still embeds prejudice in public perception. Raghav: Linguistic ambiguity doesn't reduce harm—if people interpret it as against Dalits generally, it is dangerous. Shalini: And here, "jahannum bhejo" is an explicit call for harm, making it unacceptable regardless of intent nuance. <p>Critic: So your judgment is based on both wording and likely audience perception?</p> <p>Round 3</p> <ol style="list-style-type: none"> Savita: Yes, because language in social contexts shapes real-world actions. Raghav: Especially in caste-sensitive environments, violent phrasing is inherently inciting. Shalini: Therefore, this qualifies as unambiguous hate speech.
<p>Final Classification Label - Hatespeech</p>

Figure 9: English-translated example output from GYAAN-SAHIT illustrating the full pipeline: persona assignment, multi-agent debate, critic evaluation, and final label. The original Hindi output is shown in Figure 8.