

The Fine Line Between Love & Hate

Jaee Ponde

Computer Science & Mathematics
jaee.ponde_ug25@ashoka.edu.in

Roshni Agarwal

Computer Science & Mathematics
roshni.agarwal_ug25@ashoka.edu.in

Professor Lipika Dey

Secondary Advisor
lipika.dey@ashoka.edu.in

Professor Subhashis Banerjee

Primary Advisor
suban@ashoka.edu.in

Abstract—They say there is a fine line between love and hate—but can a classifier detect it? This project examines hate speech directed at five Indian regional communities, focusing on the linguistic and cultural nuances that shape hostile expression. We constructed a manually labelled dataset and fine-tuned Indic MURIL to evaluate its ability to detect subtle, context-dependent hate in Reddit posts. Its performance was benchmarked against the base model and alternative classifiers to assess true gains from fine-tuning. To test deployability, we evaluated robustness under distribution shift using a Twitter dataset and measured variation across labels from independent annotators, revealing the deep subjectivity embedded in hate-speech judgments. Topic analyses further uncovered recurring themes and gendered patterns in community-specific hate. Ultimately, our findings show that hate-speech detection cannot rely on universal models: meaningful reliability requires context-aware, culturally grounded classifiers rather than a one-size-fits-all approach.

Index Terms—Hate speech, Indic languages, Fine-tuning, MURIL, Distribution shift, Topic modelling

CONTENTS

I)	Introduction	2
II)	Literature Review - Background & Motivation	2
III)	Project Goals & Understanding	3
IV)	Data - Selection and Scraping	3
	IV.A) Target Communities	3
	IV.B) Data Source: Reddit	3
	IV.C) Subreddit Selection	4
	IV.D) Keyword Construction	4
	IV.E) Scraping Procedure	4
V)	Labelling	4
	V.A) Methodology	4
	V.B) Authors' Labeling	4
	V.C) Data Splits	5
	V.D) Blind Evaluator Labeling	5
VI)	Model Selection	5
	VI.A) Performance Overview (Pilot Evaluation) ..	5
VII)	Fine Tuning	5
	VII.A) Dataset Used for Fine-Tuning	5
	VII.B) Rationale for Model Selection	6
	VII.C) Fine-Tuning Process	6
VIII)	Model Evaluation and Benchmarking	6
	VIII.A) Performance on Hate Class	6
	VIII.A.a) Overall Behaviour of the Finetuned Model	6
	VIII.A.b) Comparison to Other Models ..	6
	VIII.B) Performance on Not-Hate Class	7
	VIII.B.a) Overall Behaviour of the Finetuned Model	7
	VIII.B.b) Comparison to Other Models ..	7
IX)	VIII.C) Key Inferences	7
	IX) Robustness Checking	7
	IX.A) Performance on Hate Class	7
	IX.B) Performance on Not-Hate Class:	8
	IX.C) Key Inferences:	8
X)	Distribution Shifts	8
	X.A) Annotator Agreement with Ground Truth ..	8
	X.B) Annotator Agreement with Model Predictions	8
	X.C) Coincidental Alignment	9
	X.C.a) Implications	9
XI)	XI) Does the Perfect Classifier Exist?	9
	XI.A) A System Cannot be Simultaneously Comprehensive and Conservative	9
	XI.B) The Challenge of Robustness	9
	XI.C) The Inherent Subjectivity of Human Labelling	9
	XI.D) The Essential Ethical Balance: Undetected Hate vs. Over-Moderation	9
XII)	XII) Conclusion and Future Work	10
XIII)	XIII) Bonus Section : Hate Landscape (NLP)	10
	XIII.A) Text Preparation	10
	XIII.B) Sentiment Analysis	10
	XIII.C) Topic Analysis	11
	XIII.D) Gendered Patterns in Community-Directed Hate Speech	11
	XIII.E) How Regional Identities Co-Appear in Online Discourse	12
	XIII.F) Closing Thoughts	12
	References	12

I. INTRODUCTION

With the rapid expansion of social media, detecting hate speech has become more critical than ever. Effective hate detection provides guardrails against the misuse of online platforms and helps mitigate the psychological harm caused by harmful content. At first glance, deploying a well-trained classifier may seem like an obvious solution. Indeed, hate-speech classification is far from new: a wide range of approaches—from traditional machine-learning models to sophisticated transformer-based architectures—has been proposed over the years.

However, the gap between publicly available classifiers and the reliability required for real-world deployment remains substantial. This paper aims to examine that gap. We evaluate several widely used classifiers and explore what one might realistically need when deploying such systems in practice. Our focus is on issues of safety, trust, and reliability, rather than on model novelty alone.

Hate speech is inherently nuanced and highly contextual. It is subjective by nature, and achieving a fully “objective” definition is neither realistic nor necessarily meaningful, because hate is produced and interpreted by humans who understand it differently. It also varies across identities-regions, genders, age groups, communities, and socio-cultural backgrounds. Terms that are harmless within one community may be deeply offensive in another; linguistic variation further complicates detection. Moreover, hate is often conveyed through tone, context, and underlying prejudice rather than explicit slurs or keywords. It is therefore imperative to recognize that hate does not have a singular definition and never will.

To address this, we frame our task using a transparent, objective evaluation scheme that we developed for this study. We scraped and manually labelled our own dataset to fine-tune the models, ensuring that all classifiers are evaluated using the same ground truth. This prevents any model from benefiting unfairly from prior exposure to similar labelling frameworks.

Beyond classification performance, this paper also explores the diversity of hate across communities. Many forms of hate stem from deep-rooted stereotypes, and different groups experience recurring patterns of hostility. We analyse the dominant topics of community-specific hate, investigate whether geographical proximity shapes these patterns, and examine the presence of gendered biases within hate-speech content.

All our code and labelled data points can be found in our [GitHub Repository](#).

II. LITERATURE REVIEW - BACKGROUND & MOTIVATION

Automated hate-speech detection has undergone a rapid methodological evolution, beginning with lexicon-based systems and classical machine-learning classifiers before shifting toward deep learning and transformer-based architectures. Although transformer models typically achieve good performance on curated, in-domain benchmarks, a broad body of comparative research shows that these performance gains do

not generalize robustly across real-world settings [1], [2]. Studies evaluating hate-speech systems across different platforms, such as Reddit, Twitter, Facebook, and region-specific online communities show substantial drops in accuracy when models encounter unfamiliar linguistic distributions, shifting vocabularies, or new demographic contexts. **These findings suggest that many models achieve high scores by exploiting dataset-specific artifacts rather than by learning generalizable indicators of hateful intent.**

A central and widely recognized challenge concerns the detection of **implicit or subtle hate**, which rarely contains explicit slurs. Implicit hate can manifest through sarcasm, metaphor, coded expressions, community-specific euphemisms, and pragmatic insinuation. **Because these forms of hate lack obvious surface signals, they routinely bypass traditional keyword-based filters and remain difficult even for sophisticated neural architectures to recognise.** Research consistently shows that transformer models trained on standard datasets (most of which oversample explicit hate) struggle on forms of subtle, coded, or sarcastic hostility unless specifically trained on such examples [2]. Evidence from rationale-annotated resources, such as HateXplain, suggests that integrating human rationales significantly improves interpretability and the handling of subtle cases, further highlighting the limitations of datasets that emphasize explicit hate and provide insufficient linguistic context [3].

At the heart of these challenges lie foundational issues in dataset construction and annotation. **Surveys of multilingual and multicultural hate-speech corpora reveal a high degree of inconsistency in sampling strategies (e.g., keyword filtering, hashtag scraping, platform-specific crawling), annotation guidelines, and annotator training procedures** [1], [4]. The demographic background and cultural understanding of annotators have been shown to influence labelling outcomes, directly shaping the bias profile of trained models. These issues have raised concerns about label provenance and label uncertainty, prompting calls for datasets that incorporate annotator metadata, multiple annotations per instance, and explicit modelling of disagreement [5], [6] .

Dataset bias, annotation subjectivity, and narrow benchmark design have motivated a growing number of diagnostic and stress-testing studies meant to expose the brittleness of current systems. One of the most influential contributions in this direction is HateCheck, a manually constructed functional test suite designed to evaluate models on 29 specific types of hateful and non-hateful expressions. Results across numerous systems reveal systematic failures: inability to handle negation (e.g., “I don’t hate X”), inability to distinguish reclaimed slurs from targeted hate, over-reliance on profanity as a proxy for hate, and misclassification of abusive language that lacks explicit identity targeting [6]. **These findings demonstrate that high overall accuracy can mask deep structural weaknesses that would be unacceptable in a deployed setting.**

Complementary work in cross-dataset evaluation reinforces these concerns. **When models trained on one domain are evaluated on out-of-domain corpora—such as testing a Twitter-trained model on Reddit, or a dataset from one country’s linguistic community on another—performance**

often collapses, with recall showing the sharpest declines [2]. Attempts to mitigate these problems through domain adaptation, such as continued pretraining on platform-specific corpora or contrastive/multi-task domain-alignment strategies, yield only partial improvements. The overarching conclusion is that cross-platform generalization remains an unresolved research challenge, one that scaling model size alone cannot overcome.

Perhaps the clearest insight emerging from the literature is that **current models can fail quite badly under realistic conditions**. Diagnostic studies demonstrate that many systems fail to detect even simple paraphrases of hateful expressions, misclassify benign identity references as hate (false positives), and overlook coded hostility targeted at specific demographic groups (false negatives) [7]. **The consequences of these failures extend beyond academic evaluation: false negatives allow harmful content to circulate, while false positives disproportionately suppress speech from marginalized groups and cause adverse psychological effects on the poster -raising serious concerns about fairness, accountability, and the social legitimacy of automated moderation systems.**

Synthesizing across the literature, several connected gaps remain largely unresolved.

- 1) Implicit and subtle hate is inadequately represented in major datasets and insufficiently captured by standard models.
- 2) There is limited evaluation across community-level linguistic variations, such as state-level or culturally specific sub-communities within a shared language space (an especially relevant issue in multilingual and multicultural regions such as India).
- 3) Cross-platform robustness is a critical test for real-world deployment, and remains understudied despite abundant evidence of model degradation.
- 4) Many existing benchmarks provide minimal information about annotator provenance, demographic bias, or uncertainty, limiting the transparency and interpretability of model performance claims.

These gaps justify and motivate our capstone study. By constructing a provenance-aware dataset centered on subtle and community-specific hate, fine-tuning transformer models under a unified evaluation framework, and conducting cross-source robustness tests (Reddit → Twitter), this work contributes to the broader effort to build hate-speech detection systems that are contextually grounded, transparent, and empirically reliable.

In addition, we evaluate the models trained on our annotation scheme against labels produced independently by five blind participants, allowing us to quantify the inherent subjectivity in hate-speech judgements and to expose the accuracy compromises that arise when a model trained on one interpretation of hate is applied to another.

Further analyses into the topic structure of community-specific hate and the role of gender biases deepen this understanding, supporting a shift toward modelling practices that prioritize reliability, contextual awareness, and ethical sensitivity over raw benchmark performance.

III. PROJECT GOALS & UNDERSTANDING

The initial aim of this project was straightforward: to build a high-performing hate-speech classifier for Indian regional communities. However, as the project unfolded-especially during the manual labeling process-it became clear that the ambition of constructing a “perfect” hate classifier was fundamentally unrealistic. Hate is nuanced, context-dependent, culturally embedded, and inherently subjective, even under objectively defined labeling frameworks. Through this realization, the project transitioned from merely *building* a classifier to *interrogating* what it means for such a classifier to work reliably in real-world conditions.

Over the semester, these are the goals we finally aimed to achieve :

- 1) Build a labeled dataset using a clear and robust hate-speech framework.
- 2) Explore Indian regional hate, a classically understudied form of identity-based hostility.
- 3) Evaluate the trade-offs of fine-tuning, specifically how precision, recall, and accuracy shift, and whether fine-tuning meaningfully improves model performance or simply overfits.
- 4) Assess deployability, not only through accuracy - but through distribution-shift testing, robustness checks, and annotation variability.
- 5) Ask the million dollar question : **can the perfect hate classifier ever exist?**
- 6) *Bonus* : Conduct a community-wise study of Indian regional hate, exploring thematic and linguistic differences across groups.

IV. DATA - SELECTION AND SCRAPING

We collected Reddit posts that contained community-specific references and potential hate-related expressions to build a dataset reflecting geographically varied Indian online discourse. Our goal was to sample posts mentioning five target communities across different regions of India in a balanced and context-sensitive manner.

A. Target Communities

To capture India’s regional and cultural diversity, we selected the following communities, each representing a major geographic region:

- **Bihar** - Central India
- **Bengal** - Eastern India
- **Gujarat** - Western India
- **Kaarnataka** - Southern India
- **Haryana** - Northern India

These groups were chosen due to their frequent presence in online discussions related to migration, cultural stereotypes, and regional prejudice. Their strong online presence and regional distribution ensured diverse linguistic and contextual representation.

B. Data Source: Reddit

Reddit was chosen because of its:

- Open-access structure and ethical API availability through PRAW.
- High volume of India-centric discussions, memes, and commentary.
- Mix of formal and informal communication styles suitable for sentiment and hate-speech analysis.

C. Subreddit Selection

We targeted popular India-related subreddits that capture real conversational tone rather than purely informational content. Our selection criteria is reflected in the table below.

Subreddit	Users / Type
r/india	2.0M - General / National
r/indiameme	1.5M - Memes / Humor
r/Indian-DankMemes	Very high engagement - Dark / Edgy Memes
r/ABCDesis	100k+ - Diaspora / Cross-cultural
r/indiaSpeaks	Active - Politics / Opinions
r/indianpeopletwitter	Medium - Screenshots / Humor
r/indianmemes	10k–100k - Light Humor
r/IndiaSocial	Mid-sized - Social Discussion

D. Keyword Construction

We built community-specific keyword lists to capture both neutral mentions and explicitly hateful contexts.

For example, Bihari community keywords: “bihari”, “biharis”, “from bihar”, “bihar migrant”, “bihar people”, “typical bihari”, “bihari woman”, “bihari girl”, “bihari man”, “bihari boys”, “bihari AND (hate OR abuse OR insult OR rape OR harass OR attack OR violence)”

Rationale:

- Base terms to capture direct references.
- Gendered terms for gender-specific hate.
- Contextual terms to surface explicit abuse and violence discussions.

Equivalent lists were constructed for Bengalis, Gujaratis, Kannadigas, and Haryanvis.

E. Scraping Procedure

- 1) Using the PRAW Reddit API:
- 2) Iterate through each subreddit × keyword combination.
- 3) Retrieve the 100 most recent posts sorted by “new”.
- 4) Filter out deleted, empty, or removed posts.
- 5) Remove posts longer than 300 tokens to ensure model compatibility.

- 6) Store each valid entry with metadata.
- 7) Each post was stored in a structured format:

The scraping output was as follows : Gujrati | 451, Kannadiga | 301, Bihari | 272, Haryanvi | 221, Bengali | 528 **Total | 1773**

V. LABELLING

To evaluate hate speech as objectively as possible, we developed a structured annotation framework tailored to our problem: **detecting and analysing the nuances of hate toward specific Indian communities**. The framework is designed to capture not only explicit hate but also subtle expressions such as sarcasm, layered biases, contextual hostility, and comparative differences in how hate manifests across communities.

A. Methodology

Annotators were given around 25 Reddit posts per community (five communities in total) and instructed to label each post using one of three categories: Hate, Hate-Indirect, or Not Hate. The definitions were as follows:

1. Hate : A post is labeled Hate if it directly attacks, demeans, or expresses hostility toward a person or group because of their identity (community, ethnicity, language, region). This includes:

- Pejorative or discriminatory language
- Advocacy or endorsement of hostility, discrimination, or violence
- Explicit references to violence (murder, rape, physical abuse)
- Sarcasm or coded expressions targeting a group
- First-person hate expressed by the poster

2. Hate-Indirect : A post is labeled Hate-Indirect when it contains hate speech as defined above, but the hateful expression is not the poster’s direct opinion. Examples include:

- Recounting incidents involving hate
- Quoting or paraphrasing others’ hateful comments
- Narrating experiences where others expressed hate

3. Not Hate : A post is labeled Not Hate when there is no hostility or stereotype tied to community identity

- Negativity is unrelated to identity (e.g., “she is annoying because she likes pink”)
- The content is neutral, personal, descriptive, or identity-irrelevant
- This prevents general insults from being misclassified as hate speech.

In our final model training, we combined Hate and Indirect-Hate into one category. We saw that giving them as different labels during labelling ensured that evaluators actively looked for subtle hate rather than it being ignored.

B. Authors’ Labeling

All posts were manually labeled by both authors in a two-stage process:

- Independent blind labeling using the framework described above.

- Adjudication stage where disagreements were resolved through discussion.

The goal was to maximize objectivity, consistency, and conceptual alignment with our evaluation framework.

C. Data Splits

To fine-tune and evaluate our classifier, the dataset was divided into three components:

1) Fine-Tuning Data

200 posts (94 labeled as Hate, 106 as Not Hate) Used to fine-tune the Indic MURIL mode

2) Test Data

121 posts (54 Hate, 67 Not Hate) Used to evaluate the model's performance on held-out Reddit data.

3) Robustness Evaluation Data (Twitter)

49 posts (26 Hate, 23 Not Hate) - Collected from publicly available Kaggle Twitter datasets [8] from non Indian data to ensure strong distribution shift. The purpose of this dataset was to test the model's generalizability under domain mismatch (platform, linguistic register, cultural context).

D. Blind Evaluator Labeling

To examine the subjectivity inherent in hate-speech detection, we conducted an additional annotation round for the entire test dataset using five blind evaluators.

- Evaluators were aged 19–23, all current or former Ashoka University students.
- They represented three academic disciplines and five ethnic backgrounds, allowing for diverse interpretive perspectives.
- Annotators were instructed not to communicate with one another or refer to our labels.
- They were given only the labeling framework, ensuring controlled comparison.

This process allowed us to quantify how much disagreement arises from subjective interpretation and how classifier accuracy changes when evaluated against different human labeling schemes.

VI. MODEL SELECTION

Our model selection process was guided by two considerations:

- 1) The need for architectures capable of handling Indian linguistic variation, code-switching, and culturally embedded hate, and
- 2) The need for strong baselines that capture explicit hate reliably.

To evaluate these criteria, we tested four pretrained hate-speech detection models-RoBERTa-Hate [9], Indic MuRIL [10], DeHate [11], and HateXplain [12] a pilot dataset of 20 community-based posts (10 Bihari, 10 Bengali; evenly split between hate and non-hate). This preliminary evaluation allowed us to identify which models were most aligned with our goals before fine-tuning.

Model	Source	Language	Motivation
RoBERTa-Hate	Facebook / HF	English, Indian English	Strong at explicit hate; reliable baseline transformer
Indic MuRIL	Google / HF	Multilingual Indian languages	Designed for Indian linguistic contexts; handles code-switching
DeHate	CN-ERG	Indian English	Conservative classifier; context-aware but undersensitive
HateXplain	CN-ERG	Generic English	Explanation-driven model; low sensitivity but interpretable

A. Performance Overview (Pilot Evaluation)

- RoBERTa Hate - 12/20 correct
Strong on explicit hate
High confidence in predictions
Struggles with coded, sarcastic, or regionally embedded hostility
- Indic MuRIL - 16/20 correct
Performs well on multilingual or code-mixed posts
Often labels subtle or in-group sarcasm as Not Hate
Shows the highest potential once fine-tuned, given its cultural and linguistic alignment with Indian data
- DeHate - 11/20 correct
Extremely conservative; over-predicts Non-Hate
High false-negative rate
- HateXplain - 10/20 correct
No better than a random guesser

Based on the above results, we chose **Indic MuRIL for fine-tuning** because it is trained on Indian languages [10], handles code-switching, aligns culturally with regional discourse, and showed strong yet improvable baseline performance. **RoBERTa was used as a benchmark** since it is a strong English transformer that excels at explicit hate but struggles with subtle, community-coded expressions.

VII. FINE TUNING

A. Dataset Used for Fine-Tuning

The fine-tuning was performed on a custom dataset consisting of 200 data points. This dataset was carefully

curated and labeled to represent the abusive and non-abusive language relevant to the project's domain. Specifically, **94 data points were classified as ‘hate’ speech, while the remaining 106 data points were classified as ‘non-hate’ speech**. This proportional representation aimed to provide a balanced learning experience for the model, enabling it to distinguish effectively between the two classes.

B. Rationale for Model Selection

Multiple attempts were made to fine-tune the model, experimenting with different configurations, hyperparameters, and potentially varied dataset preprocessing. Through this iterative process, the current fine-tuned model was selected as the best-performing one based on its evaluation metrics (primarily F1-score) on the validation set. Other fine-tuning attempts yielded suboptimal results, demonstrating lower F1-scores, poorer generalization, or susceptibility to overfitting. The chosen model demonstrated the most robust and accurate performance in classifying abusive language within the given constraints and dataset.

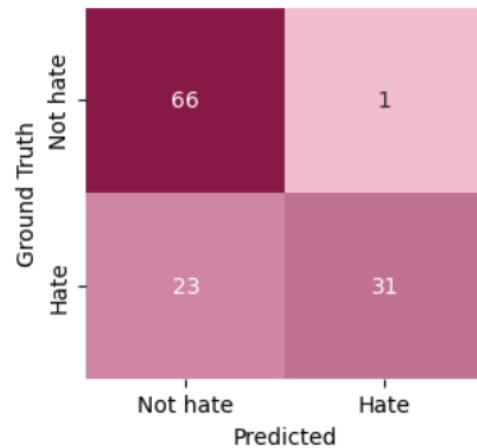
C. Fine-Tuning Process

The **Hate-speech-CNERG/indic-abusive-allInOne-MuRIL** model was fine-tuned using the Hugging Face Transformers library. The process involved training the pre-trained model on a custom dataset to adapt its performance to the specific task of identifying abusive language relevant to the project's scope. Key steps included:

- Tokenization:** The text data was tokenized using the AutoTokenizer associated with the Hate-speech-CNERG/indic-abusive-allInOne-MuRIL model, ensuring consistent input formatting.
- Dataset Preparation:** The dataset was loaded and then split into training and evaluation sets. The tokenize function was applied to encode the text, and the dataset was formatted to PyTorch tensors.
- Model Training:** The Trainer API from Hugging Face was utilized for training. TrainingArguments were configured, specifying parameters such as learning rate (2e-5), batch size (8), and number of epochs (2, with early stopping). The model was trained to minimize the classification loss.
- Evaluation:** During training, the model's performance was evaluated on a held-out test set using metrics like accuracy, precision, recall, and F1-score. Early stopping was implemented with a patience of 1 epoch, monitoring the F1-score to prevent overfitting and select the best performing model.

VIII. MODEL EVALUATION AND BENCHMARKING

To evaluate the in-domain performance of our finetuned Reddit hate-speech classifier, **we manually labelled 121 Reddit comments** and computed standard classification metrics. Because the model was trained on Reddit, this dataset allows us to assess its expected, domain-aligned behaviour, and to compare it against two baselines: Indic MURIL and RoBERTa.



Confusion Matrix of the Finetuned Indic Model

A. Performance on Hate Class

a) Overall Behaviour of the Finetuned Model:

- The precision for hate is extremely high :** Whenever the model predicts a comment as hate, it is almost always correct. The model essentially refuses to label something as hate unless it sees very strong textual evidence.
- The recall for hate is moderate :** The model fails to detect a significant portion of the actual hateful comments. A moderate recall implies that many subtly hateful or context-dependent hateful comments do not trigger the model strongly enough.
- The F1 score for hate reflects this imbalance between precision and recall :** The high precision lifts the F1, but the recall holds it back from being truly high. The balanced score suggests good performance when hate is explicit, but inconsistent behaviour for borderline or nuanced hate.
- Hate-class accuracy follows the same pattern as recall :** This metric essentially captures how many actual hate examples were correctly identified. Because recall is moderate, hate accuracy is also not very high, reinforcing that the model struggles primarily with under-detection rather than over-detection.

Model	Finetuned Indic	Indic MURIL	RoBERTa
Accuracy	0.57	0.5	0.19
Precision	0.97	0.93	0.71
Recall	0.57	0.5	0.19
F1 Score	0.72	0.65	0.29

Model Benchmarking on Hate Class

b) Comparison to Other Models:

- Compared to Indic MURIL, the finetuned model improves both precision and recall.** Indic MURIL also exhibits high precision and low recall, but the finetuned model strikes a slightly better balance, catching more hateful cases without compromising precision.
- Compared to RoBERTa, the finetuned model has dramatically better precision and recall.** RoBERTa misses most hateful comments and is less confident

even when it predicts hate. This large gap indicates that domain-specific finetuning substantially boosts performance.

B. Performance on Not-Hate Class

a) Overall Behaviour of the Finetuned Model:

- The recall for non-hate is extremely high :** The model correctly identifies nearly all non-hate comments. Very few non-hateful comments are mistaken as hate (false positives are rare). **This aligns with the model's clear tendency to lean toward “non-hate” unless hate is unmistakable.**
- The precision for non-hate is moderate.** Although the model predicts “non-hate” frequently, some of those predictions are incorrect - i.e., some true hate comments get classified as non-hate. **This is the mirror image of the hate-class behaviour: instead of wrongly flagging innocent comments, the model more often “forgives” hateful content.**
- The F1 score reflects this behaviour : excellent recall with meaningfully lower precision.** This results in a strong but not exceptional F1, indicating reliability but also that the model can over-generalize the non-hate label.
- Non-hate accuracy is extremely high :** This reflects the dominance of correct predictions for non-hate - again because recall is nearly perfect. **The model is extremely good at protecting non-hateful content from being misclassified.**

Model	Finetuned Indic	Indic MURIL	RoBERTa
Precision	0.74	0.71	0.59
Recall	0.99	0.97	0.94
F1 Score	0.85	0.82	0.72
Accuracy	0.99	0.97	0.94

Model Benchmarking on Not-Hate Class

b) Comparison to Other Models:

- The finetuned model improves upon Indic MURIL across all non-hate metrics.** This improvement signals that the finetuning process made the classifier much more stable and consistent for everyday, non-hateful language.
- The model also outperforms RoBERTa, especially in precision and F1.** RoBERTa tends to confuse edge cases and mislabels more content than the finetuned model.

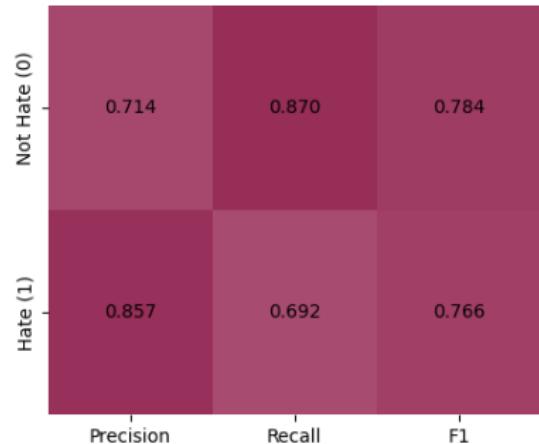
C. Key Inferences

- The model is precision-oriented, especially for hate.** It almost never produces false alarms. But this comes at the cost of missing real hate instances.
- There is a **systematic asymmetry between the two classes:**
 - Hate: high precision, lower recall - cautious and selective.
 - Non-hate: very high recall, moderate precision - lenient and forgiving.

- Together, these behaviours indicate a foundational bias toward predicting non-hate unless hate is very explicit.
- The dominant error type is false negatives in the hate class.** This means actual hateful comments get classified as non-hate more often than the other way around.
- Compared to Indic MURIL and RoBERTa, the finetuned model is clearly superior.** Finetuning makes the model more precise, more stable, and better aligned to the Reddit domain. The gains are especially evident when dealing with ambiguous content.

IX. ROBUSTNESS CHECKING

To assess the robustness and generalizability of the finetuned Reddit hate-speech model, we evaluated it on 49 manually labelled Twitter posts. Unlike Reddit, Twitter was not part of the training data, making this an out-of-domain test. This evaluation helps reveal how well the model copes with new linguistic environments - including shorter text, heavy use of slang, hashtags, emojis, and platform-specific discourse patterns.



Performance of Finetuned Indic model on Twitter data

A. Performance on Hate Class

- While the precision for the hate class drops compared to Reddit, it still remains high :** Even on Twitter, when the model labels content as hate, it is usually correct. Despite domain shift, the classifier still avoids making reckless or overly aggressive hate predictions. This suggests the core decision boundary for hate remains stable.
- The recall for hate improves compared to its in-domain behaviour :** The model captures a larger proportion of actual hateful tweets than it does for Reddit comments. This is particularly interesting given the unfamiliar platform: Twitter content seems to trigger the hate classifier more readily. The improvement in recall implies that Twitter hate-language, often being more direct or condensed, may be easier for the model to detect.
- The F1 is higher on Twitter,** indicating that the model has moved closer to equilibrium between precision and recall. The domain shift effectively causes the model to

be less conservative and more willing to assign the hate label.

- **Accuracy for the hate class aligns with this pattern :** The classifier correctly identifies a higher proportion of actual hateful tweets than hateful Reddit comments. The types of hateful expressions in Twitter samples lie closer to the decision-patterns the model learned during training.

B. Performance on Not-Hate Class:

- **Recall for non-hate is high, although lower than on Reddit :** The model still identifies a large majority of non-hate tweets correctly. However, the drop from its near-perfect in-domain recall indicates it is slightly more cautious on Twitter, predicting hate more often than before. **This change is consistent with the model being less conservative and more responsive to hate-like cues in the unfamiliar domain.**
- **Precision for non-hate is moderate.** Some tweets predicted as non-hate are actually hateful - these are false negatives. This implies the model still struggles with borderline hate cases, but now in a different environment where slang and irony may obscure intent.
- **The F1 score for the non-hate class remains strong.** It indicates that despite domain shift, the model retains a consistent ability to recognise safe content. The F1 also reflects that although recall remains strong, precision has room for improvement.
- **Non-hate accuracy is reasonably high.** The classifier correctly labels most non-hateful tweets, though not as overwhelmingly as on Reddit. This suggests that the shift in domain introduces ambiguity or linguistic noise that the model does not completely understand.

C. Key Inferences:

- Domain shift effects are clear: **The model is more willing to label content as hate on Twitter.** This reflects a shift toward higher recall and slightly lower precision across both classes.
- Hate detection improves out-of-distribution: Surprisingly, the model catches more actual hate on Twitter than on Reddit. **The nature of Twitter hate - short, direct, and often explicit - seems to better match what the model has learned.**
- Non-hate predictions weaken slightly: The model still performs well but loses some of the extreme confidence it had in-domain. It now generates more false positives for hate, indicating increased sensitivity to hate-like cues.
- **Primary error remains false negatives in hate, though at a lower rate than in-distribution.** The model still misses some hate, but fewer than on Reddit.

X. DISTRIBUTION SHIFTS

Hate-speech classification is not a purely objective task. Even when multiple annotators receive the same definition

and guidelines, their interpretations may differ because hate is contextual, culturally situated, and often ambiguous [5], [7].

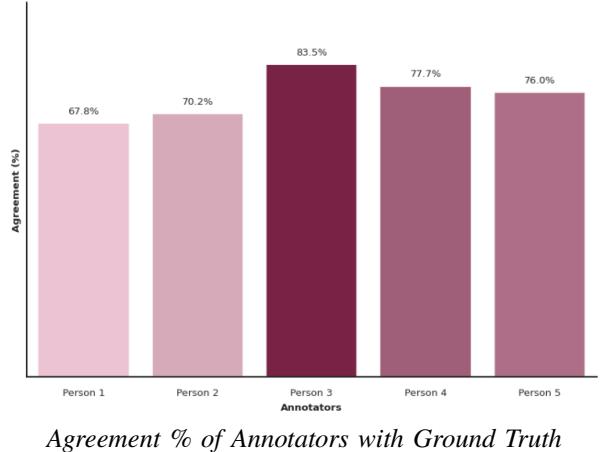
This analysis examines how five independent annotators labelled the same 121 Reddit posts and compares their labels to:

- Ground truth (labelling of the primary annotators - Jaee and Roshni)
- Model predictions

A. Annotator Agreement with Ground Truth

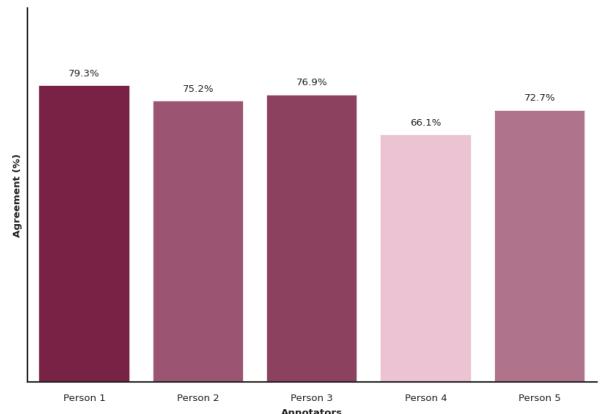
Even with identical instructions and definitions, annotators display substantial variability: Agreement ranges approximately from **68% to 84%**. The highest-agreement annotator aligns somewhat closely with the ground truth. The lowest-agreement annotator deviates by approx 16 percentage points from the top performer. This means:

- Human annotation is inherently noisy. Even with the same definition, annotators interpret tone, intent, and context differently.
- Hate detection is nuanced. Ambiguities-sarcasm, coded language, indirect slurs, intra-group usage-lead to legitimate disagreement among humans.
- Ground truth is not absolute. It reflects one “expert” interpretation, but not necessarily unanimous consensus.



Agreement % of Annotators with Ground Truth

B. Annotator Agreement with Model Predictions



Agreement % of Annotators with Model Prediction (Overall Accuracy)

When comparing annotators' labels to the model's predictions, an interesting pattern emerges:

- **Person 1, who had the lowest agreement with the ground truth (68%), shows the highest agreement with the model (79%).**
- In contrast, annotators who were closer to the ground truth show slightly lower agreement with the model.

C. Coincidental Alignment

- **Person 1 labelled only 15 out of 121 datapoints as hate, a much lower proportion than the ground truth.**
- The model tends to under-predict hate, likely due to limited examples of subtle hate in the training set or conservative thresholds learned from the ground truth.
- The high agreement between Person 1 and the model does not indicate the model "adapting" to the annotator, since the model was trained only on the ground truth.
- **Instead, the sparse hate labelling by Person 1 coincidentally matches the model's tendency to miss certain hateful examples.**

a) Implications:

- Model conservatism: The model struggles with nuanced or implicit hate. This tendency aligns with Person 1's extremely cautious labelling.
- Variance in annotator style amplifies perceived distribution shift: Different annotators' label distributions highlight how metrics can vary substantially depending on who labels the data.
- **High agreement with a conservative annotator does not necessarily indicate better performance - it may reflect a shared bias toward under-predicting hate.**

XI. DOES THE PERFECT CLASSIFIER EXIST?

Benchmarking the fine-tuned model revealed a **fundamental, structural trade-off in classifier design that makes perfection a statistical impossibility**.

A. A System Cannot be Simultaneously Comprehensive and Conservative

- **The High-Recall Problem (Over-Detection):** A model optimized for maximal recall is designed to catch most of the hate, ensuring very little harmful content is missed. However, this strategy generates an unacceptable number of false positives. It will invariably misclassify satire, counterspeech, or powerful instances of reclaimed language as hate, leading to the suppression of legitimate voices.
- **The High-Precision Problem (Under-Detection):** Conversely, a model optimized for maximal precision aims for high confidence, ensuring that whatever it flags as hate is almost certainly correct. Like our fine-tuned system, this approach leaves a significant amount of harmful content undetected (false negatives). This is particularly evident in new domains like Twitter, where the model's conservative nature fails to generalize.

This insight shows that model metrics like precision and recall are not universal scores for certification. Instead, robust evaluation must treat these metrics as situated and platform-specific, intrinsically tied to the specific linguistic environment they are tested in.

B. The Challenge of Robustness

- Our analysis underscores that **true deployability hinges on robustness**—a classifier's ability to maintain performance when deployed on data with a different distribution than its training set.
- The success of fine-tuning was clear: it improved performance and gave a more faithful representation of Indian regional hate on Reddit, the platform of its origin. **However, when tested on Twitter, the model showed its vulnerability to distributional differences.**
- For real-world application, this pattern suggests the fine-tuned system should be deployed as a high-precision filter rather than a comprehensive detector.
- It is excellent for reliably flagging a subset of clearly hateful content and confidently discarding most benign posts, but its conservative behavior means it will systematically underestimate hate on other platforms like Twitter.

C. The Inherent Subjectivity of Human Labelling

Even if a model could magically generalize across platforms, it faces a more fundamental human obstacle: the deep subjectivity embedded in hate-speech judgments. Hate speech is inherently nuanced, culturally embedded, and subjective. The variability in the labelling of the five independent blind evaluators showed us:

- **Human annotation is inherently noisy, even with a structured framework.** Ambiguities like sarcasm, coded language, indirect slurs, and context lead to legitimate disagreement among humans.
- Even within our group of five annotators—all from the same university, similar age range, and comparable educational backgrounds their labels varied significantly. **This shows us that interpretation of hate is highly prone to the labeller's bias.** Demographic factors, personal experience, and individual sensitivity to sarcasm or coded language shaped judgments in different ways, demonstrating that **disagreement can arise even in seemingly homogeneous groups.**
- The ground truth is not absolute; it reflects only one expert interpretation, not a universal consensus. A model trained on one set of human judgments will experience accuracy compromises when evaluated against a different, yet equally valid, human interpretation.

D. The Essential Ethical Balance: Undetected Hate vs. Over-Moderation

The failure of the "perfect" classifier is, at its core, an ethical dilemma. Any real-world deployment must navigate a crucial and delicate balance:

- **The Risk of Undetected Hate:** When hate speech or targeted harassment goes unflagged, users from the

affected communities experience heightened stress, anxiety, and a sense of being unsafe or unwelcome. Repeated exposure to hostility also normalises harmful norms for bystanders, reinforcing prejudice and contributing to a feedback loop where toxicity becomes part of the platform’s identity.

- **The Risk of Over-Moderation:** For many users, especially those whose cultural expressions, dialects, or political rhetoric fall outside the linguistic patterns seen in training data, over-moderation produces a constant sense of surveillance and precarity. **This disproportionately affects marginalised groups, who often rely on vernacular, reclaimed slurs, humour, or coded language that classifiers routinely misinterpret.**

In light of our final, million-dollar question, **the inferences suggest that a truly perfect hate classifier is unlikely to exist**, because of both : model limitations and because *hate* itself is contextually and politically contested.

The takeaway for our project is not that fine-tuning has failed - it clearly improves performance and gives a more faithful representation of Indian regional hate on Reddit.

A “good” model evaluation must treat metrics as platform specific and intrinsically tied to bias, rather than as universal scores that could ever certify a hate classifier as perfect or fully deployable across all contexts.

XII. CONCLUSION AND FUTURE WORK

This project demonstrates that a perfect hate-speech classifier—one capable of capturing every nuance, cultural signal, and subjective interpretation—does not and cannot exist. **Instead, future work must shift toward understanding what the closest viable, reliable, and ethically deployable alternative looks like, and how such systems should be designed, evaluated, and monitored in real-world settings.** This reframing turns the problem from “How do we build a flawless model?” to “How do we build a responsible one?”

To advance this direction, several foundational questions emerge:

- 1) If a perfect classifier is unattainable, what is the closest functional and ethically acceptable model we can design?
- 2) What questions should we ask before deploying a hate-speech classifier in any real platform or community context?
- 3) How can we measure trust and reliability beyond accuracy (particularly across annotators, demographic groups, and distribution shifts)?
- 4) What do model word embeddings reveal about latent biases, stereotyping, or learned associations that may shape predictions?
- 5) Should hate speech models always be context-specific regionally, linguistically, or culturally tailor?
- 6) What do model word embeddings reveal about latent biases, stereotyping, or learned associations that may shape predictions?

What constitutes a robust evaluation framework, and what components (e.g., interpretability, robustness checks, annotator disagreement, cross-domain testing) must it include?

Taken together, these questions form a broader thesis: the future of hate-speech detection lies not in producing universal models, but in building culturally grounded, and context aware systems. Progress in this field will depend on frameworks that recognize linguistic diversity, human subjectivity, and the social impact of automated moderation moving us toward models that are not perfect, but responsible.

XIII. BONUS SECTION : HATE LANDSCAPE (NLP)

Our NLP component was designed to understand the patterns underlying hate directed toward Indian regional communities. Because our labeling framework already captured subtle and indirect hate, we used NLP primarily to explore how hate is expressed across communities, and to identify recurring themes, tones, and contextual markers that might escape simple keyword-based analysis.

For the NLP section, we ran all non-training scraped data points through the fine tuned indic muril and dropped those that exceeded the NLP models token length. This resulted in total posts : 1322, Out of which 326 were Hate. We treated all posts labeled “Hate” by the classifier as containing hate speech for the purpose of downstream linguistic analysis.

A. Text Preparation

To prepare text for analysis, we applied minimal preprocessing-removing URLs and formatting while preserving emojis, slang, spelling variations, and code-switching, since these elements often signal sarcasm or implicit bias.

B. Sentiment Analysis

We applied VADER (Valence Aware Dictionary and Senti-ment Reasoner), a rule-based sentiment model optimized for social-media text.

We used adjusted sentiment thresholds to better capture the emotional nuances of Reddit posts. Posts with a compound score of 0.5 or higher were classified as positive, any score at or below 0 was labeled negative, and Scores between 0 and 0.5 were treated as neutral These thresholds were chosen to better reflect the intensity and ambiguity typical of community-based online discourse.



Sentiment analysis of all 5 communities

Across all five communities, the sentiment distributions show that discussions involving regional identity on Reddit are **rarely neutral and tend to be emotionally charged**. While **negative sentiment is substantial-especially for Haryanvi and Bihari posts**-positive sentiment is also prominent across communities, suggesting that these conversations include cultural pride, humor, or supportive commentary alongside hostility.

Neutral sentiment remains consistently low, indicating that community references generally evoke strong reactions. Overall, the patterns illustrate that online discourse about Indian regional groups is mixed rather than uniformly negative, with both positive and negative sentiments shaping how these identities are talked about.

C. Topic Analysis

We used LDA-based topic modeling to identify the dominant themes within posts for each community. After preprocessing the text, we built a token dictionary and filtered out extremely rare or overly common words to focus on meaningful patterns. Using this cleaned corpus, we trained a Latent Dirichlet Allocation (LDA) model-automatically adjusting the number of topics when data was sparse-to extract clusters of frequently co-occurring terms.

Community	Topic	Representative Keywords
Bengali	Topic 1	hindu, state, west, election, person, movie, language, hindi, bjp, post
	Topic 2	movie, genocide, bangladeshi, illegal, party, assam, pakistan, muslim
Bihari	Topic 1	leader, migrant, crisis, come, train, bengali, school, punjab
	Topic 2	family, culture, temple, name, party, year, take, political, state
Gujarati	Topic 1	bjp, wedding, delhi, christian, muslim, custom, language
	Topic 2	case, test, hindi, positive, rate, city, american
Haryanvi	Topic 1	government, beef, crore, bjp, hindu, journalist, report
	Topic 2	hindu, report, violence, use, woman, may, protest, thing, life, edit
Kannadiga	Topic 1	south, comment, communist, indian, rumor, hateful, thodi
	Topic 2	city, hindi, language, taminadu, state, english, party, part, capital

Top 2 dominant topics of all 5 communities

Bengali :

- The top Bengali topics revolve around religion, state politics, and immigration narratives, with strong references to “Bangladeshi,” “illegal,” “Assam,” “Pakistan,” and “genocide.”
- These keywords reflect how **Bengali identity online is often entangled with debates around cross-border**

migration, NRC/Assam issues, and communal politics.

- Overall, Bengali discourse is heavily politicized and tied to national-security narratives rather than purely cultural identity.

Bihari :

- Bihari topics foreground migration, labour mobility, and education, with keywords like “migrant,” “train,” “school,” and “punjab.”-
- Bihari narratives oscillate between migrant stereotypes and cultural-political identity**, reflecting how Biharis are frequently framed through socio-economic status.

Gujarati :

- Topic 1 blends politics (BJP) with religion (Christian, Muslim) and cultural customs, highlighting how Gujarati identity is often associated with political allegiance and cultural conservatism.
- Topic 2 introduces the **striking presence of the keyword “american,” which likely appears due to references to the “Gujarati model” or diaspora-related commentary**. This suggests a narrative where Gujarati identity is connected to diaspora success, making it both locally and globally framed.

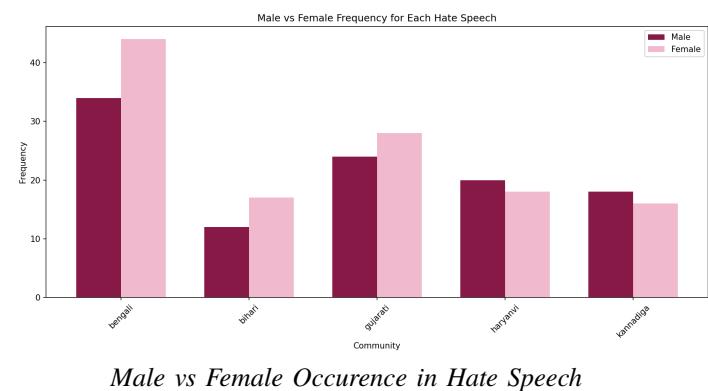
Haryanvi :

- Haryanvi topics prominently feature beef politics, government, protests, violence, and “journalist,” pointing to a discourse shaped by **media controversies and communal tensions**.
- The presence of “woman,” “protest,” and “violence” in Topic 2 indicates that gendered issues and public demonstrations play a role in how Haryanvi identity is discussed.

Kannadiga

- Kannadiga topics focus heavily on language politics, evident from “hindi,” “language,” “tamilnadu,” “english,” and “capital.” Topic 1 also includes “communist,” “rumor,” “hateful,” suggesting that online discourse around Kannadigas intersects with inter-state linguistic tension.
- The Hindi-vs-South linguistic dynamic is clearly visible. **Overall, Kannadiga topics emphasize regional linguistic pride and political positioning, especially in relation to North–South debates.**

D. Gendered Patterns in Community-Directed Hate Speech



This analysis compares how often male-coded and female-coded keywords appear in hate-classified posts for each community, offering insight into whether hate is gendered.

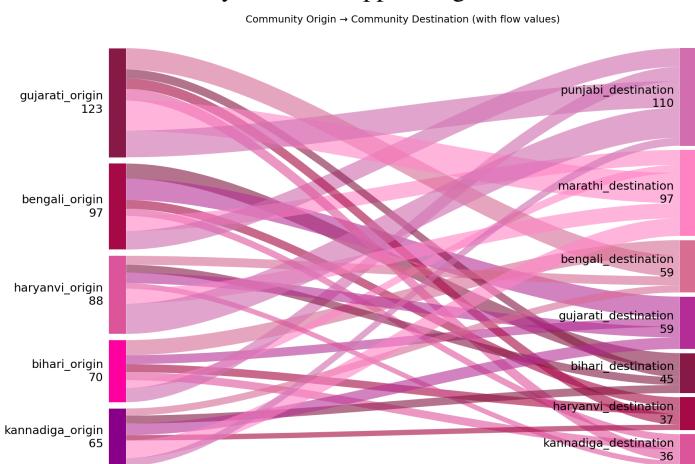
- Across most communities, female-coded terms occur slightly more frequently than male-coded ones, with the **Bengali subset showing the largest gap-indicating that hate towards Bengalis may disproportionately reference or target women.**
- Gujarati and Bihari posts show a similar pattern, where female mentions exceed male mentions, suggesting gendered framing in community stereotypes.
- In contrast, Haryanvi and Kannadiga posts show a more balanced distribution, with only small differences between male and female references.

These trends suggest that gender plays a meaningful but community-specific role in how hate is expressed, with some groups—especially Bengalis-linked to more female-targeted narratives.

E. How Regional Identities Co-Appear in Online Discourse

To explore how different regional communities are referenced together within the same post, we constructed a set of community-specific keyword lists (e.g., Bengali, Gujarati, Bihari, Kannadiga, Haryanvi).

For each post, we created binary indicator columns showing whether keywords from each community appeared in the text. Using these indicators, **we computed pairwise co-occurrences, identifying cases where a post scraped for one community (the origin) also contained keywords referring to another community (the destination).** These cross-references were then visualized using a Sankey diagram, where the width of each flow represents the frequency with which two community identities appear together in the dataset.



Community based co-occurrence in hate speech

- Punjabi and Marathi references appear widely across origins : **Large flows toward punjabi_destination and marathi_destination suggest that these identities enter discussions even when the original post focuses on another group.** This reflects their cultural prominence

and frequent appearance in national-level conversations, memes, and stereotypes.

- Strong cross-links between linguistically or geographically adjacent communities such as Gujarati → Marathi, Haryanvi → Punjabi, Bihari → Bengali. **These align with known migration corridors, interstate tensions, or linguistic overlaps.**
- Kannadiga shows fewer but more topic-specific references : Flows from Kannadiga origin tend to be narrower, reflecting that hate or discussions involving South Indian identity are more linguistically contained and less frequently entangled with North-Indian community stereotypes.
- Cross-community hate is not isolated : This demonstrates that **hate (or general discussion) about one community often invokes stereotypes, comparisons, or slurs about others.** Hate expressions online are therefore part of a networked discourse, not standalone community attacks.

F. Closing Thoughts

Taken together, this section highlight that hate toward Indian regional communities is deeply shaped by political narratives, migration histories, linguistic tensions, and gendered stereotypes.

These patterns reveal that hate speech is not a uniform linguistic phenomenon but a community-specific one, where meaning depends heavily on cultural context and social identity. **As a result, universal hate speech detectors trained on generic or Western datasets are fundamentally limited in their ability to capture the subtleties of Indian online discourse.**

Community-based studies like ours are therefore essential: they uncover the local dynamics that models routinely miss, provide culturally grounded data for fine-tuning, and expose the representational gaps that lead to misclassification.

A fuller understanding of hate speech, and the responsible deployment of models that detect it, requires attending to these community-level nuances rather than assuming a single shared linguistic landscape.

REFERENCES

- [1] T. Chakrabarty, A. Mittal, and others, “NBias: A Natural Language Processing Framework for Bias Identification in Text,” *Neurocomputing*, 2023, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231223003557>
- [2] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, “Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus,” *Language Resources and Evaluation*, 2019, [Online]. Available: <https://www.researchgate.net/publication/331433247>
- [3] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection,” *arXiv preprint arXiv:2202.09517*, 2022, [Online]. Available: <https://arxiv.org/abs/2202.09517>
- [4] D. Yang, X. Li, and S. Kumar, “Understanding Implicit and Subtle Hate Speech in Online Communities,” *IEEE International Conference on Big Data*, 2024, [Online]. Available: <https://ieeexplore.ieee.org/document/10848067>
- [5] Z. Waseem, “Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection,” in *Proceedings of NAACL-HLT*, 2016, p. ii–x. [Online]. Available: <https://aclanthology.org/N16-2013>

- [6] L. Zhong, K. Shu, and H. Liu, “A Survey on Cross-Domain and Cross-Lingual Hate Speech Detection,” *ACM Transactions on Knowledge Discovery from Data*, 2020, [Online]. Available: <https://arxiv.org/abs/2004.08527>
- [7] D. Kumar, K. Garimella, and others, “Revisiting the Labeling of Offensive Language: Annotation Schemes and Annotator Demographics Matter,” in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2021, p. –. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/18063>
- [8] vkrahul, “Twitter Hate Speech Dataset.” [Online]. Available: <https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech>
- [9] D. Antypas and J. Camacho-Collados, “Twitter-RoBERTa-base for Hate Speech Detection.” [Online]. Available: <https://huggingface.co/cardiffnlp/twitter-roberta-base-hate>
- [10] M. Das, S. Banerjee, and A. Mukherjee, “Data Bootstrapping Approaches to Improve Low Resource Abusive Language Detection for Indic Languages,” *arXiv preprint arXiv:2204.12543*, 2022, [Online]. Available: <https://arxiv.org/abs/2204.12543>
- [11] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, “Deep Learning Models for Multilingual Hate Speech Detection,” *arXiv preprint arXiv:2004.06465*, 2020, [Online]. Available: <https://arxiv.org/abs/2004.06465>
- [12] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection,” *arXiv preprint arXiv:2012.10289*, 2020, [Online]. Available: <https://arxiv.org/abs/2012.10289>