JOURNAL OF COMMUNICATION, LANGUAGE AND CULTURE

Detecting Hate Speech in Hindi Digital Discourse Using Transformer–Long Short-Term Memory Models

Rachna Narula[1], Vedika Gupta[2*], Jawad Khan[3], Tariq Rahim[4], Vijay Kumar[1]

[1]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[2]Jindal Global Business School, OP Jindal Global University, Haryana, India
[3]School of Computing, Gachon University, Seongnam, Republic of Korea
[4]Department of Computer Science, Kingston University, London, United Kingdom

*Corresponding author: vgupta2@jgu.edu.in; ORCID iD: 0000-0002-8109-498X

ABSTRACT

Hate speech on digital communication platforms has become a major obstacle to healthy online discourse, especially in multilingual societies such as India, where Hindi is a dominant language in social media interactions. Hostile, offensive, and defamatory speech is linguistically and socio-culturally complex because of colloquial idioms, regional variations, code-mixing, and culturally embedded references. However, effective detection is crucial for creating safer and more inclusive digital communication environments. This study evaluates advanced language models for analysing hate speech in Hindi social media content. A dataset of 21000 Hindi posts from Twitter and public repositories, categorized into five categories, was analysed: hate, offensive, fake, defamation, and non-hostile. General-purpose models were tested against Hindi-specific language models, including Hindi-BERT (based on Bidirectional Encoder Representations from Transformers, or BERT) and MuRIL, to investigate whether performance can be further enhanced by integrating Long Short-Term Memory (LSTM) layers. In tests, the best-performing model, Hindi-BERT (with sequential learning added), correctly identified 94 out of every 100 posts—a considerable improvement over simpler models. For social media platforms, it has real-world consequences: it can automatically identify harmful content to be reviewed by a human, minimize nuisance alerts that waste human moderators' time, and identify defamatory or bogus posts before they gain a lot of traction. The results offer a systematic approach to researching how hostility develops in the Hindi-speaking online community, how linguistic creativity (e.g., slang, sarcasm, code-mixing) can conceal or manifest hostility, and how decisions about content moderation influence public discourse for communication scholars. Overall, this paper illustrates that language-specific computational tools can be used for both platform governance and communication research, provided that the cultural context is considered. Finally, technical methods are combined with communication scholarship to explain and curb harmful speech in the online public sphere.

Keywords: digital communication, Hindi language, hate speech detection, online discourse analysis, social media communication, transformer-based language models, cultural and linguistic analysis, content moderation

Received: 28 January 2026, **Accepted:** 23 April 2026, **Published:** 30 July 2026

1.0 Introduction

For many years, social media has provided a forum for developing and exchanging ideas (Alshalan and Al-Khalifa, 2020). Although this provides users the opportunity to express their opinions, feelings, and emotions, some people exploit these platforms to create and spread offensive and malicious information (V. Gupta et al., 2021). Today, hate speech refers to content directed toward an individual, group, community, or nation. The lack of suitable datasets in various languages, including Hindi (Khezzar et al., 2023), hampers research and development. Hate speech is harmful content that explicitly incites or supports hatred against individuals (Mubarak et al., 2020; Nath et al., 2026) or groups based on race, religion, caste, age, gender, sexual orientation, or ethnicity (Almaliki et al., 2023). Hate speech on social media refers to statements or posts that defame, smear the reputation of, or insult a group of people or individuals within that group (Pawar et al., 2022).

The line between hate speech and other types of offensive speech is a fine one, and these nuances have important practical implications for detection systems. Hindi-speaking communities are spread throughout the world and regularly use Hindi as the main language platform on social media, reinforcing the significance of the proper identification of hate speech in this population (Bansod, 2023). Hindi is among the most widely spoken languages in the world, with approximately 615 million speakers worldwide, ranking third among the world's languages (Mossie & Wang, 2018). The linguistic nature of Hindi, including its orthographic conventions, syntax, and code-mixing, makes identifying hate speech even harder (Bashar & Nayak, 2020). Accordingly, a thorough understanding of such linguistic intricacies is indispensable for developing effective countermeasures against online hate speech (Bohra et al., 2018).

Hate speech in digital communication is not simply a problem of categorising words and expressions but a serious threat to the use of language, cultural expressions, and social interaction in the digital environment (Papacharissi, 2016; Tufekci, 2014). Hindi, as a medium, is widely used across multiple digital media, such as social media, news commentaries, public debates, and discussions, in environments dominated by multiple languages, like India (Sharma et al., 2023, 2024, 2025; Nath et al., 2025). The use of hostile, defamatory, and offensive language in Hindi often mirrors the socio-cultural tensions, power dynamics, and ideological conflicts in the larger socio-cultural landscape of the society (Udupa, 2018).

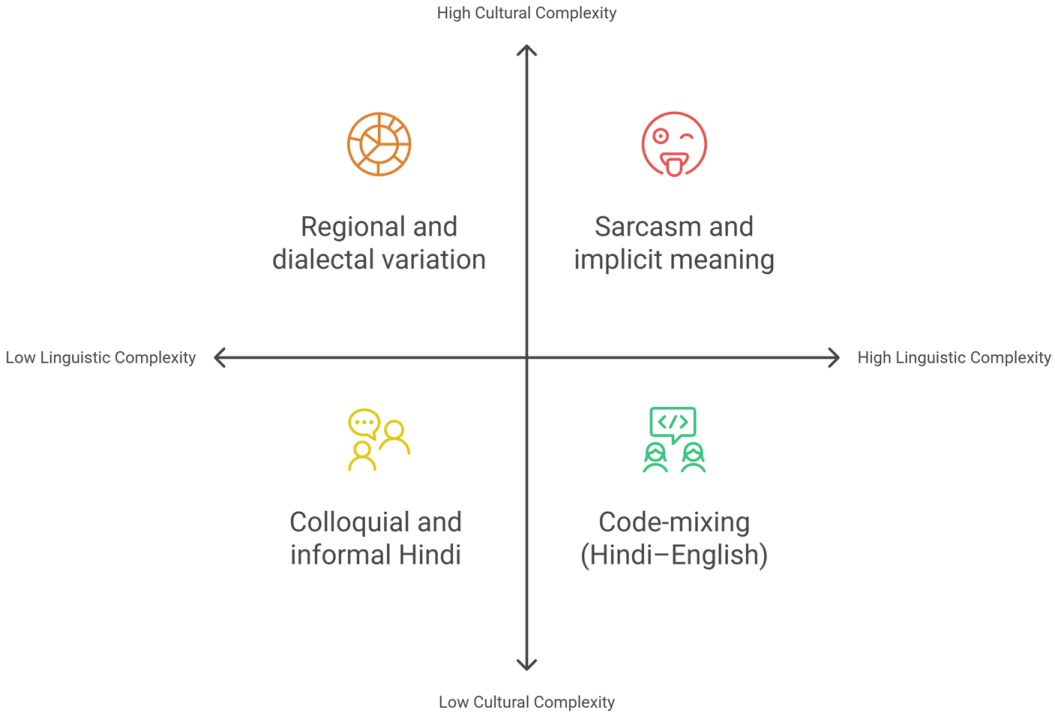
The rapid expansion of negative-quality language on social media normalizes the use of aggressive language and polarizes societies (Garg et al., 2024; Mathew et al., 2019; Zannettou et al., 2020). In informal written Hindi, slang, regional variations, code-mixing, and culturally embedded words and phrases are part of everyday language, making hate speech detection (Bohra et al., 2018; Bansal et al., 2022). These linguistic qualities are problematic not only for computational approaches but also for researchers seeking to understand how digital communication works and the cultural narratives it carries (Fortuna & Nunes, 2018).

In terms of communication and linguistics, automated hate speech detection may help foster a healthier online communication environment by allowing quick moderation and providing content to help inform policy decisions (Gillespie, 2019; Mondal et al., 2017). Computational approaches provide systematic tools for analysing large amounts of user-generated text and can generate insights into patterns of recurring hostile behaviour and how they spread through digital communities (Schmidt & Wiegand, 2017). This investigation centres on material in Hindi and connects technological innovation with broader communication practices in culturally diverse online contexts (Bansal et al., 2022). Accurately recognizing and classifying hate speech in Hindi online spaces is essential to prevent the dissemination of harmful content and foster safer online environments (Saleem et al., 2017). As social media and internet platforms on the Indian subcontinent continue to expand, Hindi content in online communication is an increasingly significant proportion (Kemp, 2022). Having reliable categorization strategies that are sensitive to dialectal variation, slang, and cultural references is crucial in preventing offensive and discriminatory content and preserving linguistic diversity (Waseem & Hovy, 2016). Furthermore, there are nuances to be made between harmful speech and legitimate opinions, and moderation strategies seek to balance the two; in doing so, they must distinguish between them (Crawford & Gillespie, 2016).

Therefore, maintaining a balance between user safety and balanced discourse within the Hindi-speaking community is of critical importance, given the omnipresence of hate speech on online platforms and its potential effects on both individuals and the community (Bilewicz & Soral, 2020). Hindi-BERT and MuRIL are promising advanced language models that can help enhance the identification and moderation of hate speech (Sharma et al., 2025; Swamy et al., 2019). When deployed, these models can help more accurately identify and classify malicious content, enabling platforms to take proactive steps and mitigate potential negative consequences. Language-specific models, e.g., Hindi-BERT, are especially valuable because they address challenges related to linguistic nuances and cultural aspects of Hindi-speaking populations (Bansod, 2023). However, none of the models can be definitively described as the best, and further research and collaboration are needed to improve methods for detecting hate speech against new language forms and emerging hate speech online (Founta et al., 2018). Language and cultural factors affecting the production of hate speech in Hindi are illustrated in Figure 1.

Figure 1

Linguistic and cultural dimensions influencing hate speech in Hindi digital communication and their analysis through computational language models.



This study makes three contributions to research on hate speech detection in Hindi digital communication:

(i) First, it implements and evaluates a hybrid Transformer–LSTM framework that combines the contextual representation capability of transformer-based language models, namely Hindi-BERT and MuRIL, with the sequential learning strength of Long Short-Term Memory (LSTM) networks. While previous studies have predominantly employed transformer models or recurrent neural networks independently for detecting hate speech in Hindi (Swamy et al., 2019; Jadhav et al., 2021; Sharma et al., 2023), the present study investigates their integration within a unified framework for improved contextual and sequential understanding of hostile online discourse.

(ii) Second, the study constructs and evaluates a combined dataset of 21,000 Hindi social media posts by integrating 13,000 self-collected tweets with 8,000 publicly available hostility

instances. Compared with several datasets used in previous studies on Hindi hate speech, which typically relied on smaller or platform-specific collections (Swamy et al., 2019; Sharma et al., 2023), the resulting dataset provides broader linguistic and contextual coverage for model evaluation.

(iii) Third, this study presents a comparative evaluation of LLaMA-family models (TinyLLaMA, Airavata, and OpenHathi), BERT-family models (Hindi-BERT and MuRIL), and hybrid Transformer–LSTM architectures under a common experimental framework. This comparison provides benchmark evidence regarding the relative effectiveness of generative, contextual, and hybrid approaches for hate speech classification in Hindi digital communication.

Unlike prior studies that typically focus on a single model family or a single dataset, this study integrates a larger Hindi dataset, evaluates multiple transformer architectures, and investigates hybrid Transformer–LSTM models within a unified framework. This combination enables a more comprehensive assessment of methods for detecting hate speech in Hindi digital communication.

2.0 Literature Review

This section provides a thorough description of the most recent focus studies and literature reviews on the detection of hate speech in Hindi-language social media communication. Prior research has been conducted from both computational and linguistic perspectives in the field of hate speech detection using both classical machine learning models and deep learning and transformer models. These works shed light on the problems of linguistic diversity, code-mixing, and culturally embedded expressions in Hindi online discourse. The literature reviewed provides a basic understanding of the role that computational approaches can play in analysing and moderating hostile language in digital communication environments (Jhunthra et al., 2025).

Transformer architectures such as BERT (Devlin et al., 2019) and language-specific variants, including MuRIL and Hindi-BERT, have demonstrated strong capability in capturing contextual semantic information in multilingual environments. In contrast, Long Short-Term Memory (LSTM) networks remain effective for modelling sequential dependencies and long-range contextual patterns in text (Hochreiter & Schmidhuber, 1997). Hybrid Transformer–LSTM architectures seek to combine these complementary strengths, motivating their application to hate speech detection in Hindi digital communication (Vaswani et al., 2017; Devlin et al., 2019).

The studies summarized in Table 1 demonstrate a gradual shift from traditional machine learning methods to transformer-based and ensemble approaches, reflecting the growing need to capture contextual and sequential patterns in Hindi social media communications.

Although previous studies have evaluated Transformer-based models such as MuRIL, IndicBERT, and mBERT, limited attention has been given to hybrid Transformer–LSTM architectures for Hindi hate speech detection. This gap motivates the present study, which combines contextual Transformer representations with sequential learning mechanisms to improve classification performance while maintaining linguistic sensitivity to Hindi digital communication.

The F1-scores reported in Table 1 range from 0.60 to 0.91. Lower scores (e.g., de Paula et al., 2023, with 60% precision) reflect the inherent difficulty of Hindi code-mixed data, where the same word can have different connotations depending on context. Higher scores (e.g., Sharma et al., 2023, with F1 0.91) come from smaller, more focused datasets (2,020 tweets) that may not generalize broadly. Our study builds on this prior work in three ways: (1) we use a substantially larger dataset (21,000 posts); (2) we evaluate both LLaMA-family models and BERT-family models together; and (3) we systematically test ensemble architectures (Transformer + LSTM) to capture sequential dependencies often missed by pure transformer models. The wide range of prior results confirms that no single model dominates, motivating our comparative approach.

Table 1

Summary of Studies on Hate Speech Detection in Hindi

Study	Dataset	Methods	Results
Sharma et al. (2024)	TABHATE (2,020 Hindi tweets)	CNN, LSTM, Bi-LSTM, mBERT, MuRIL, IndicBERT, XLM-R	F1: 0.91
de Paula et al. (2023)	HASOC-2021, CONSTRAINT-2021	mBERT + Adam	Precision: 60%
Jadhav et al. (2021)	HASOC 2020, 2021	BERT-CNN	F1: 77%
Hadj Ameur & Aliane (2021)	HASOC 2019	CNN	Accuracy: 82%
De La Peña Sarracén (2021)	HASOC Hindi tweets	LR, NB, SVM, RF	Macro F1: 69% (A), 63% (B)
Raman et al., (2022)	HASOC 2019 and FB-aggregated data TRAC	DistilBERT	DistilBERT Acc: ~86%
Swamy et al. (2019)	HASOC Hindi	DistilBERT, MuRIL, embeddings	MuRIL+BERT best
This Study	21,000 tweets	Hindi-BERT+LSTM, MuRIL+LSTM, LLaMA family, BERT family	Accuracy: 93.75%, F1: 0.9106

3.0 Dataset

The dataset comprised 21,000 Hindi social media posts, including 13,000 tweets collected via web scraping and 8,000 instances from the publicly available Hostility dataset (Bhardwaj et al., 2020). These data are divided into five categories: non-hostile, defamatory, fake, hateful, and offensive. Several categories can be used to classify posts that share similar features. This dataset has been made publicly accessible by the collaborative project CONSTRAINT-2021, which is focused on identifying the most harmful posts. When assembling the data, the authors recognized the possibility of bias, especially in regard to the selection of the tweets, which might reflect the prejudices of society. Tweets were gathered using specific hashtags and user accounts, which may have introduced unintended bias. The direction of future work will be to reduce these biases by adding more data sources and perspectives that are more heterogeneous. Decisions relating to false tweets, which include misleading and inaccurate content, and offensive tweets, which may make use of hate speech or other derogatory expressions, will be narrowed down, as well as defamatory statements, which are destructive of the reputation of a person. The non-hateful tweets do not meet the requirements of other categories. These classification criteria aim to maintain the accuracy and uniformity of labelling across the dataset by carefully balancing linguistic factors, contextual knowledge, and expertise.

a. Hate: Posts that promote violence or hatred towards certain people due to their race, ethnicity, religion, or place of residence, as in Table 2.

b. Defamation: False information intended to damage the social reputation of a person or organization, shown in Table 2.

c. Fake: Information or claims that are proven to be false, shown in Table 2.

d. Offensive: Posts that contain profanity, vulgarity, or other offensive content directed at specific individuals or groups, shown in Table 2.

e. Non-Hostile: A non-hostile message is also known as a good tweet or a neutral tweet, shown in Table 2.

Table 2

Example Hindi tweets with English translations and classifications

Hindi Text	English Translation	Category
यहाँ के मुसलमानों को देश छोड़ देना चाहिए	"Muslims here should leave the country."	Hate
XYZ नेता एक ठग है; उसने लाखों रुपये चुराए	"Leader XYZ is a fraud; he stole millions of rupees."	Defamation
चंद्रयान-2 चाँद पर उतरा, भारत ने इतिहास रच दिया	"Chandrayaan-2 landed on the moon; India made history."	Fake (misinformation)
तुम बहुत बेवकूफ हो, कुछ भी नहीं जानते	"You are very stupid, you know nothing."	Offensive
आज मौसम बहुत सुहाना है, बाहर घूमने चलते हैं	"Today's weather is very pleasant, let's go out."	Non-Hostile

3.1 Self-built Dataset

Using the Scraper API, dataset was extracted from Twitter containing up to 100 tweets per term. The dataset was collected using terms, expressions, and topics related to hate discourse in Hindi. To enable a more thorough examination, the dataset's composition was carefully chosen to supply a reasonable representation of both non-hateful and hateful speech. This is a multi-label classification task, meaning a single post can belong to multiple categories simultaneously (e.g., a post can be both "hate" and "offensive" – see the composite labels above). This differs from multi-class classification, where each post belongs to exactly one category. The 16 composite labels emerged naturally from the co-occurrence of the five base categories (hate, offensive, fake, defamation, non-hostile) during annotation. For example, many posts containing defamatory content also used offensive language, resulting in the label "defamation, offensive."

Annotation protocol: Three native Hindi speakers (ages 24–35, all holding graduate degrees, with diverse regional backgrounds from Delhi, Lucknow, and Bhopal) independently annotated the self-built dataset. Each annotator was trained on a separate set of 200 example posts with consensus labels. The base categories were defined as:

Hate: Direct incitement to violence or hostility based on identity (race, religion, caste, gender).

a. Offensive: Profanity, vulgarity, or insulting language not necessarily targeting a protected group.

b. Defamation: False statements that damage reputation.

c. Fake: Proven false information presented as fact.

d. Non-hostile: None of the above.

The annotators assigned all applicable categories to each post. Inter-annotator agreement was calculated using Cohen's kappa (Cohen, 1960) for pairwise agreement on each of the five base categories. The

average kappa across all three annotator pairs was 0.82 (substantial agreement, Landis & Koch, 1977). Disagreements (approximately 12% of posts) were resolved by majority vote; in cases of three-way disagreement (less than 2% of posts), the post was excluded from the final dataset. These definitions were provided to annotators in written guidelines and reviewed in training sessions. To disambiguate overlapping label semantics, the following operational definitions for posts that received multiple labels were adopted:

1. Defamation + Hate: A post that falsely damages reputation AND incites violence/hatred (e.g., falsely accusing a religious group of criminal acts while calling for action against them).
2. Defamation + Offensive: A post that damages reputation using profanity or vulgar insults without inciting violence.
3. Hate + Offensive: A post that expresses hatred using profane language but does not contain false claims.
4. Defamation + Hate + Offensive: A post containing all three – false, damaging claims, incitement to hatred, and profanity.

3.2 Dataset Creation and Merging

As mentioned earlier, comments were initially extracted using web scraping from tweets (Mubarak et al., 2021). To do this, tweets are filtered based on keywords, hashtags, or user accounts, and then the resulting dataset is combined with the publicly available 'Hostility' dataset. This is an important step as it could help extend the dataset and make the model more widely applicable and generalise better. Merging the self-created dataset with the public Hostility dataset increased data diversity and improved model generalization across platforms.

Under-sampling and over-sampling techniques were employed to balance the classes (He & Garcia, 2009; Chawla et al., 2002). Oversampling boosted the proportion of hostile categories by random duplication with small changes to the features, and under-sampling shrank the proportion of non-hostile tweets by randomly discarding tweets. The bias removal strategy used in this work produced bias-free and consistent training data with better generalization performance for the model. This approach has proven effective, as demonstrated by improvements in cross-category metrics in Section 5 (Chawla et al., 2002). This balanced sampling helps mitigate skewness, which can be observed in some real-world social media datasets (Mubarak et al., 2021).

3.3 Data Preprocessing

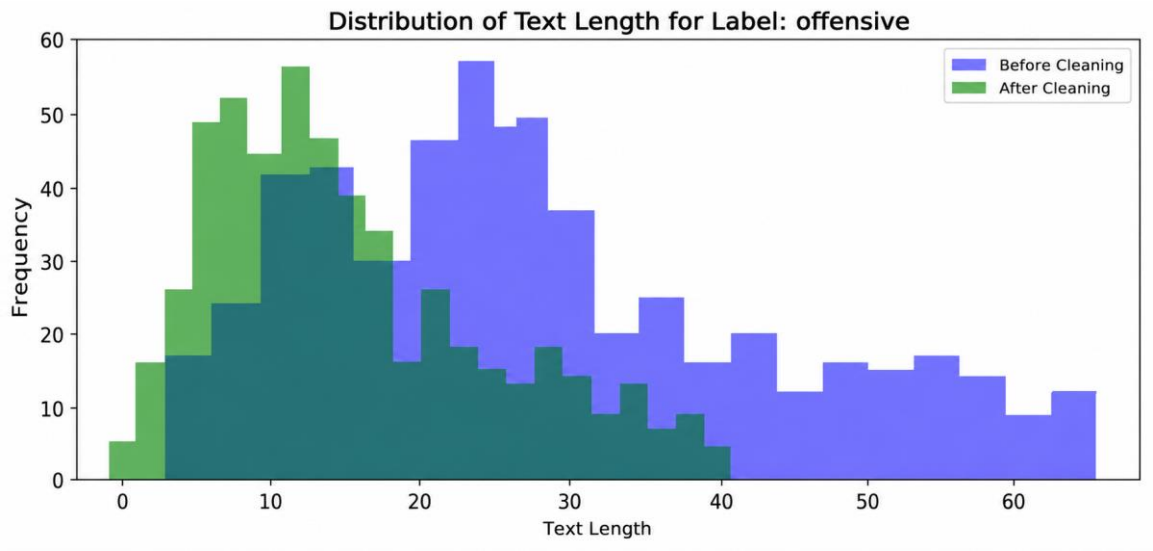
Text preprocessing followed standard practices for Hindi social media data. The URLs, non-Hindi characters, and extra whitespace were removed. Two preprocessing choices require justification because they are non-obvious:

1. Removal of emojis: Preliminary experiments on a 500-post sample showed that including emojis did not improve classification F1-score (0.71 with emojis vs. 0.72 without). Because emojis are often platform-dependent and add noise, they were removed using regular expressions.
2. Removal of English words and code-mixing: Hindi social media posts frequently contain English words. However, the goal was to evaluate Hindi-language models only. Therefore, English words (detected via Unicode Devanagari block filtering) were removed to focus the model on Hindi linguistic patterns. This discards potentially meaningful code-mixed signals; future work should examine code-mixed preservation (Bansal et al., 2022; Gupta et al., 2025). This preprocessing decision was made to isolate the performance of Hindi-specific language models rather than to model naturally occurring code-mixed communication.
3. Stop-word removal used a standard Hindi stop-word list (from the Natural Language Toolkit). Tokenization was performed using the tokenizers provided with each transformer model (BERT WordPiece for Hindi-BERT/MuRIL, LLaMA tokenizer for LLaMA models). After

preprocessing, the average tweet length decreased from 158 characters to 112 characters, largely due to the removal of URLs and English words. Figure 2 presents distribution of text length for label "offensive".

Figure 2

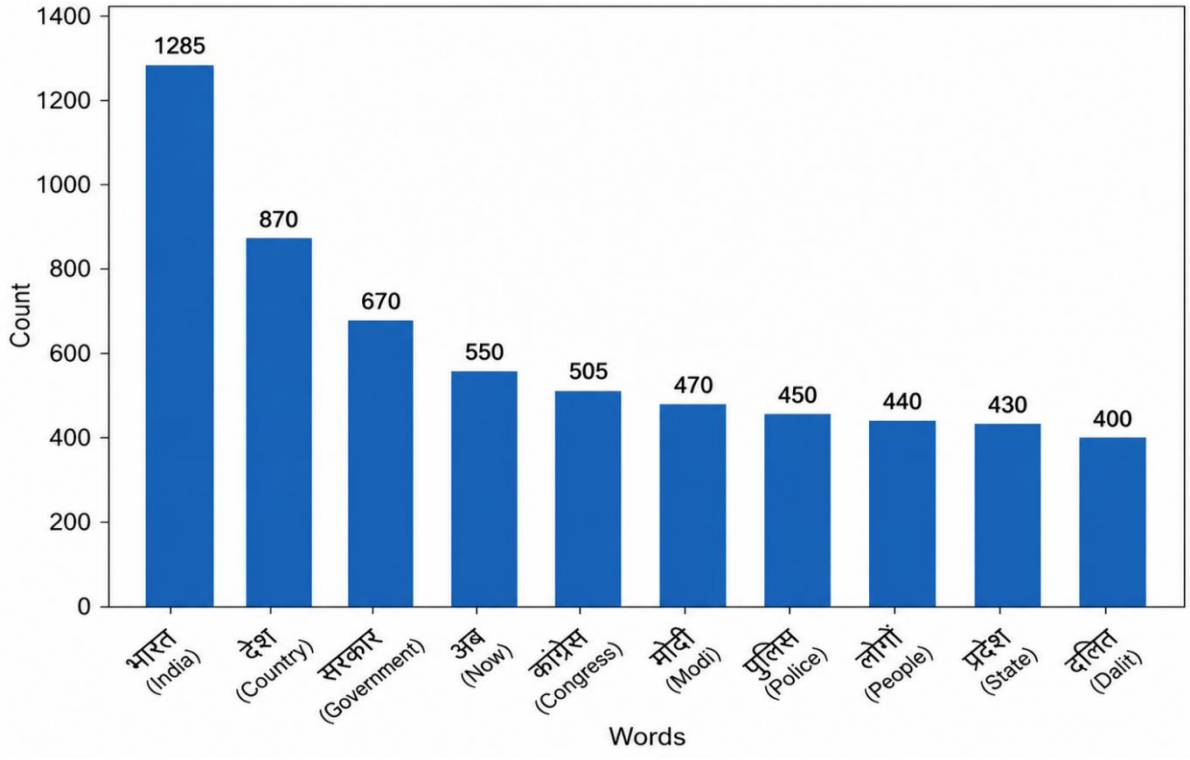
Reduction in the length of words under the label 'offensive'



Following preprocessing, it was observed that some words were more frequent than others, and that the latter were critical for sentiment and speech analysis. The plot of the Top 10 most common terms, which is shown in Figure 3, illustrates this.

Figure 3

Top 10 most frequent Hindi words with English translations.



3.4 Feature Extraction using BERT Embeddings

Contextual embeddings using the [CLS] token from Bidirectional Encoder Representations from Transformers (BERT) models, specifically Hindi-BERT and MuRIL, were extracted. The [CLS] representation captures sentence-level semantic information by aggregating contextual relationships across the entire input sequence. These embeddings were subsequently used as input features for downstream classification models, including the proposed Transformer–LSTM architectures.

4.0 Methods

4.1 Model Training & Testing

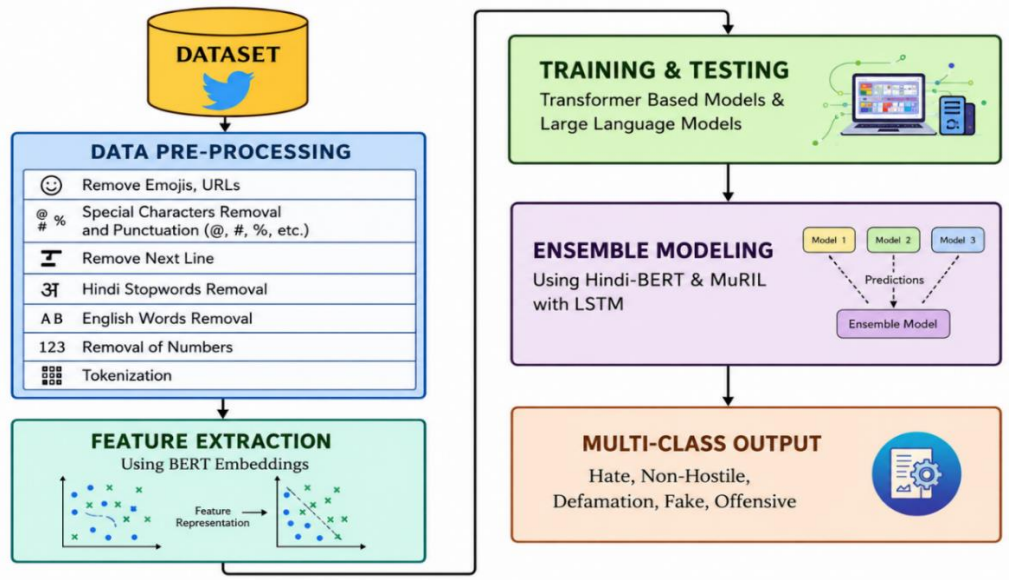
This section describes the computational methodology adopted to analyse hate speech in Hindi digital communication. The emphasis of this methodology is not on architectural novelty but on applying transformer-based and hybrid language models to capture contextual and sequential patterns in real-world Hindi social media discourse. Figure 4 illustrates the overall workflow of the proposed approach, which is explained in detail in the following subsections.

To analyse hate speech in Hindi digital communication, this study evaluates a diverse set of transformer-based and hybrid learning models. The evaluated models include LLaMA-family models (TinyLLaMA, Airavata, and OpenHathi), BERT-family models (MuRIL and Hindi-BERT), and hybrid Transformer–LSTM architectures that combine contextual transformer representations with Long Short-Term Memory (LSTM) networks. These models were selected to investigate the relative contributions of contextual semantic understanding and sequential learning mechanisms in the detection of hate speech within Hindi social media discourse.

The dataset was then split into training and validation subsets in order to ensure that the models are able to generalize to unseen data. Standard classification metrics, such as accuracy, precision, recall, and F1-score, which are commonly used in the body of research on hate speech detection, were used as model performance measures.

Figure 4

Methodology Adopted



4.2 LLaMA Models

Later in 2023, Meta AI published the LLaMA family of large language models, which were trained to solve multiple tasks of natural language processing. In this paper, models based on LLaMA were used as the comparative baselines to determine their feasibility in Hindi hate speech classification. TinyLLaMA is a small model that has around 1.1 billion parameters trained on large-scale textual data. Its comparatively small size and reduced computing power make it appropriate to experiment in resource-limited environments. OpenHathi, developed by Sarvam AI, was included as a comparative Hindi-oriented language model alongside Airavata. The two models were initially designed for areas where instructions are required but were changed in this case to the classification of hate speech. These models were fine-tuned using the Low-Rank Adaptation (LoRA) mechanism. LoRA adds trainable rank-decomposition matrices into transformer layers, which can be easily fine-tuned with lower computation cost. The adaptation also enables LLaMA-based models to perform sentiment and hate speech detection on Hindi text while preserving their representational strengths.

4.3 BERT Models

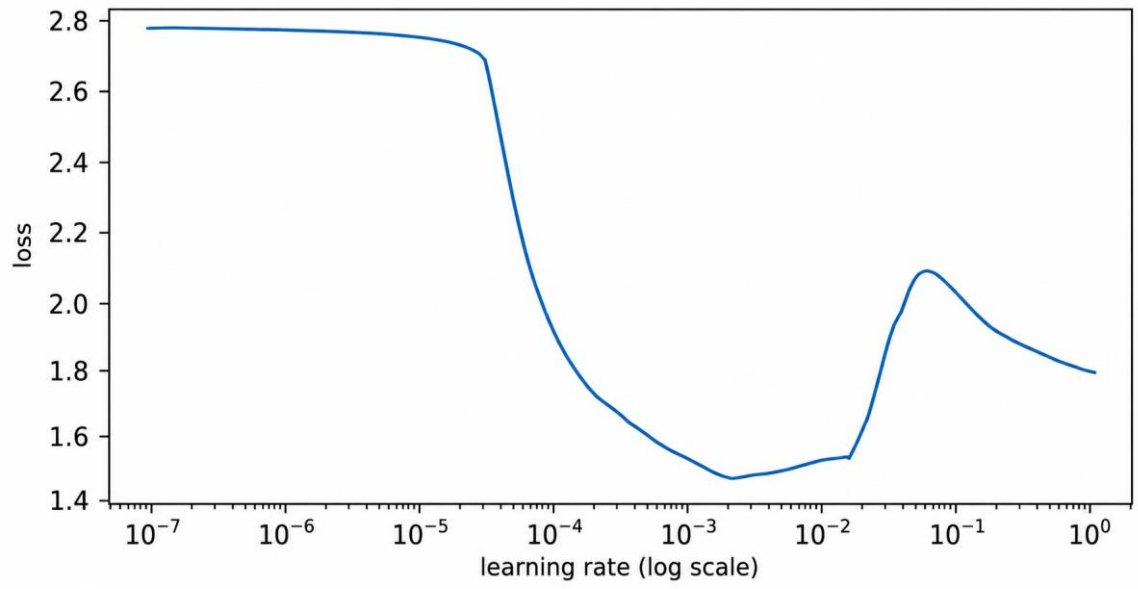
BERT models can capture contextual information in text using bidirectional word representations. Indian-specific BERT variants MuRIL and Hindi-BERT were employed in this investigation.

Transformer-based MuRIL supports various Indian languages, including Hindi. Phonetic and transliteration-based pre-training helps handle Indian language variety. MuRIL is adept at sentiment analysis and named entity identification, making it suitable for multilingual and code-mixed hate speech detection.

Hindi-BERT is a language model trained on Hindi corpora to detect Hindi text's syntactic and semantic features. Hindi-BERT represents subtle Hindi digital communication language patterns via large-scale pre-training and task-specific fine-tuning. Figure 5 shows the learning-rate-versus-loss behaviour during Hindi-BERT training.

Figure 5

Learning Rate versus Loss for Hindi-BERT Training



4.4 Ensemble Models

The principal methodological contribution of this study is the evaluation of Transformer–LSTM ensemble architectures for detecting hate speech in Hindi. To better depict Hindi social media conversations, these models use transformer-based contextual embeddings and LSTM-based sequential learning.

LSTM layers receive transformer model output embeddings (MuRIL or Hindi-BERT) in the proposed architecture. Transformer models capture contextual information via attention processes, whereas LSTM networks describe sequential relationships and maintain long-range contextual information across text sequences. Ensemble models can capture the global contextual semantics and sequential linguistic patterns of hate speech across clauses or sentences.

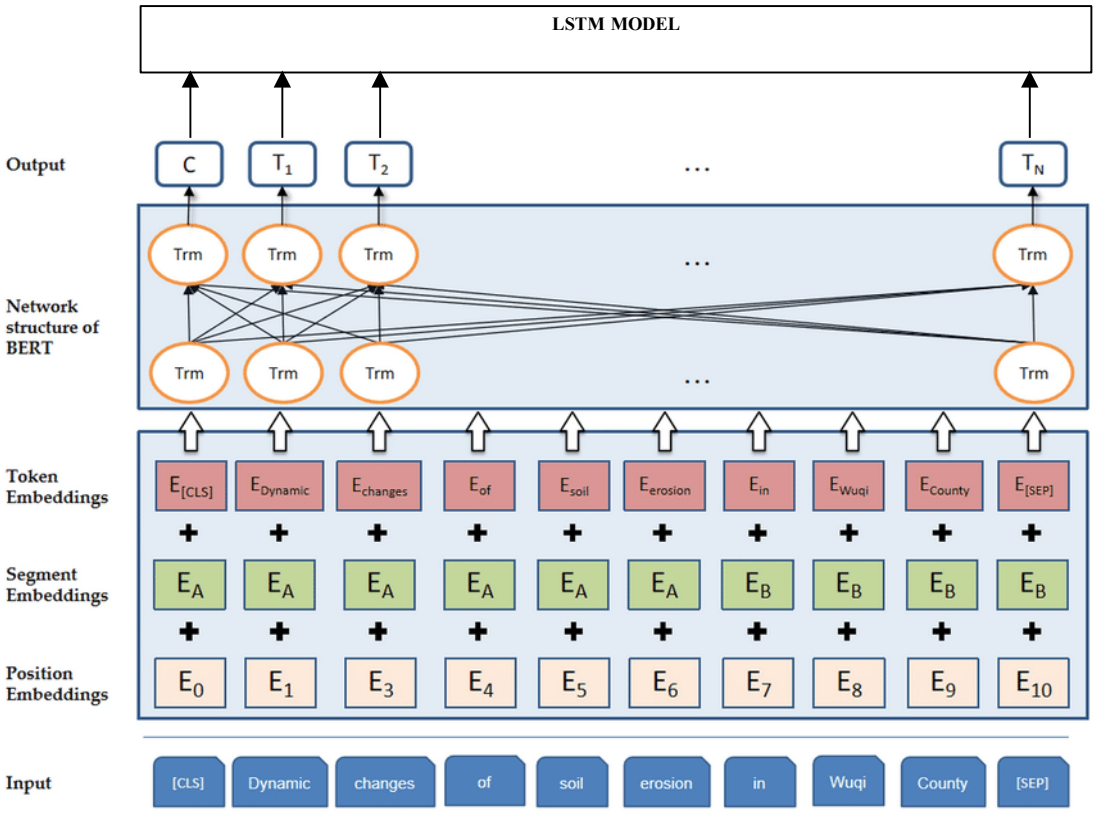
The ensemble configurations were MuRIL and Hindi-BERT with LSTM. Figure 6 depicts BERT–LSTM ensemble construction. Ensemble models learned from contextual embeddings and sequential patterns by optimizing model parameters using gradient descent and backpropagation.

Combining LSTM layers with transformer models improved the model's capacity to identify latent animosity in Hindi social media messages. This is particularly useful for recognizing oblique language, sarcasm, and prolonged hostile discourse.

In general, transformer-based contextual modelling and LSTM-based sequential learning work together to improve the model's performance. The experimental results presented in Section 5 demonstrate the effectiveness of this approach.

Figure 6

BERT with LSTM



5.0 Experimental Setup and Results

5.1 Experimental Setup

The learning rate was set to 1e-4, and the maximum sequence length was set to 256 for all experiments. The models were trained using transformer-based architectures. The GPU T4 x2 used for training the models on Kaggle had a total of 30 GB RAM. To avoid overfitting, the dataset was split into training, and validation sets in a 90:10 ratio. Several cycles of hyperparameter tuning were performed, and a learning rate of 1e-4 and a batch size of 32 were chosen to train the model. These values were optimized for grid search to obtain a trade-off between model performance and speed of convergence.

5.2 Results and Analysis

The results of the implemented models are presented here, such as LLaMA-based models (TinyLLaMA, OpenHathi, and Airavata), BERT-based models (Hindi-BERT and MuRIL), and ensemble models that were used to combine BERT variants with LSTM. In this study, the following evaluation metrics were used: Precision, Recall, and F1-score.

5.2.1 Analysis of BERT Model

MuRIL Model:

The training and evaluation of the MuRIL model were performed for 10 epochs. The model accuracy increased from 29.93% to 75.90%, and the training loss decreased from 2.740 to 1.729. The model had a high recall of 0.857, which means that the model shows stable learning behaviour and generalization performance. The final accuracy and loss values indicate that the model effectively learned from the training data. The training accuracy, loss, precision, recall, and F1-score values for MuRIL across all 10 epochs are shown in Table 3.

Hindi-BERT Model:

Hindi-BERT was trained for a total of 10 epochs with a single-cycle learning rate of 0.00012. The accuracy of the training increased from 56.42% to 82.73%, and the loss decreased from 1.58 to 0.5374. The results showed that the accuracy of the validation was 74.30% with a loss of 0.9298, indicating the capability of the model in generalizing to unseen data, as shown in Table 4.

Table 3

Training Accuracy and Loss for MuRIL for all 10 Epochs

Epoch	Acc (%)	Loss	Precision	Recall	F1
1	29.93	2.740	0.226	0.428	0.285
2	65.29	2.662	0.761	0.857	0.799
3	66.35	2.474	0.857	0.857	0.857
4	68.47	2.331	0.428	0.428	0.428
5	70.81	2.186	0.428	0.428	0.428
6	72.18	2.147	0.714	0.714	0.714
7	73.40	2.018	0.428	0.285	0.342
8	73.91	1.891	0.785	0.714	0.714
9	74.87	1.820	0.714	0.571	0.634
10	75.90	1.729	0.857	0.857	0.857

The training and validation accuracy trends for Hindi-BERT across epochs are illustrated in Figure 7, which shows consistent improvement during training.

Table 3 summarizes the training and validation accuracies and loss values for Hindi-BERT across ten epochs.

Figure 7

Accuracy v/s Epochs

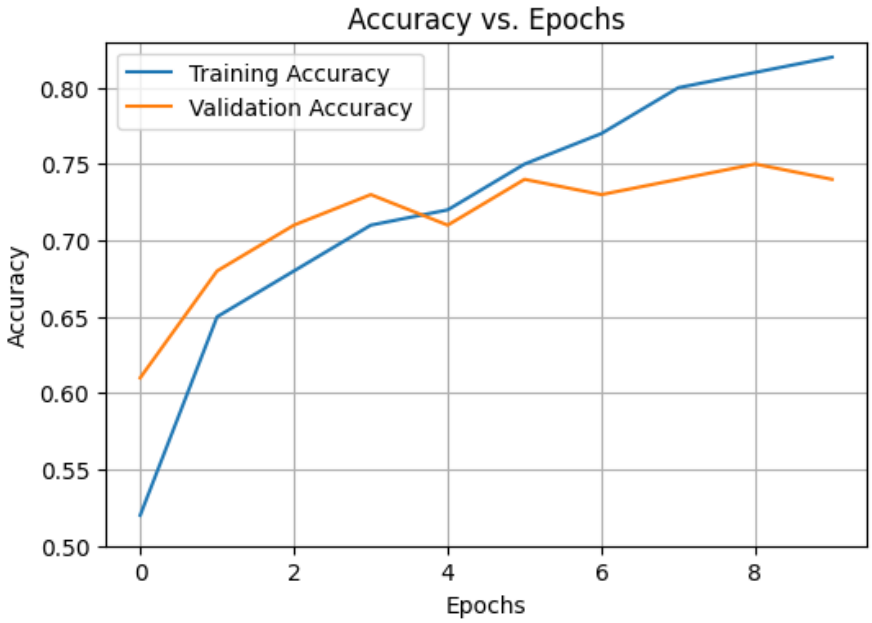


Table 4

Training and Validation: Accuracy and Loss for Hindi-BERT for all 10 Epochs

Epoch	Train Acc (%)	Train Loss	Val Acc (%)	Val Loss
1	56.42	1.5803	65.01	1.0847
2	66.71	1.0410	70.32	0.9462
3	70.41	0.9417	70.67	1.0138
4	73.56	0.8517	72.97	0.8767
5	74.42	0.7932	72.00	0.8856
6	76.28	0.7372	73.18	0.8833
7	77.63	0.6875	71.16	0.9932
8	79.70	0.6160	73.60	0.8746
9	80.87	0.5769	74.51	0.9132
10	82.73	0.5374	74.30	0.9298

5.2.2 Analysis of LLaMA Models

Tiny LLaMA achieved 56% accuracy, Airavata 59%, and OpenHathi 61.5% (Table 5). OpenHathi performed best among the LLaMA-family models, although all lagged the BERT-based models.

Table 5

Analysis of Tiny LLaMA, Airavata, and OpenHathi

Models Used	Accuracy (%)	Training Loss	Train Runtime	Total_Flos
TinyLLaMA	56	0.558	24:09.84	7914376gf
Airavata	59	0.6133	1:01:45.95	23859057gf
OpenHathi	61.5	0.6008	1:00:10.28	23866233gf

5.2.3 Analysis for Ensemble Models

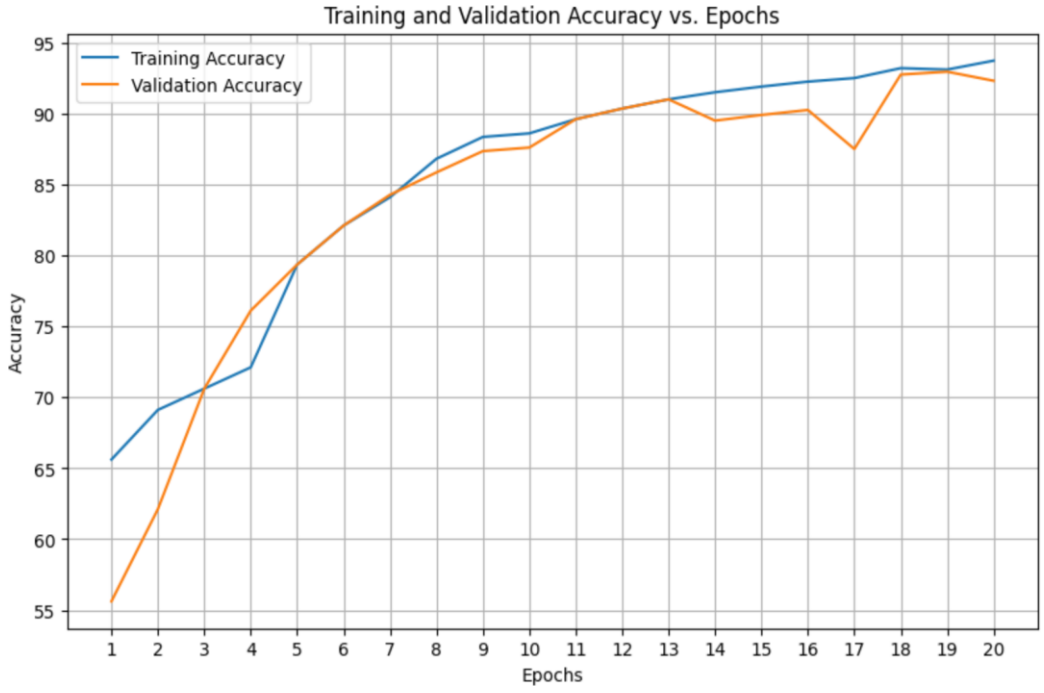
MuRIL with LSTM: The MuRIL with LSTM ensemble model was trained over 19 epochs, achieving a final accuracy of 86.16% and a training loss of 0.4528. The validation accuracy reached 75.34%, indicating improved robustness in detecting hate speech.

Hindi-BERT with LSTM**:** The Hindi-BERT with LSTM ensemble model was trained over 20 epochs and achieved the highest accuracy of 93.75% with a training loss of 0.2341. The validation accuracy was 92.33% with a validation loss of 0.3609. The model also achieved a precision of 0.8902, recall of 0.8757, and an F1-score of 0.9106.

Figure 8 illustrates the accuracy versus epochs for the Hindi-BERT with LSTM model. MuRIL+LSTM accuracy improved to 86.16% after 19 epochs. Hindi-BERT+LSTM achieved 93.75% accuracy after 20 epochs.

Figure 8

Accuracy vs Epochs graph for Hindi-BERT with LSTM



5.2.4 Transformer Based Model Comparison

The results show that Hindi-BERT with LSTM achieves the highest accuracy (93.75%) and F1-score (0.9106), outperforming all individual models. Hindi-BERT (82.73%) surpasses MuRIL (75.90% accuracy) for Hindi-specific hate speech detection, despite MuRIL's strong performance on multilingual tasks. Among LLaMA-family models, OpenHathi (61.5%) performs best, though all lag BERT-based models. The ensemble approach (transformer + LSTM) provides a clear performance gain, confirming the value of sequential learning for hate speech detection in Hindi social media text. Table 6 tabulates this comparison.

Table 6

Comparative Analysis of all Models

Model	Acc (%)	Loss	Precision	Recall	F1
TinyLLaMA	56.00	0.5580	0.56	0.80	0.72
OpenHathi	61.50	0.6008	0.59	0.83	0.74
Airavata	59.00	0.6133	0.61	0.85	0.76
MuRIL	75.90	1.7290	0.857	0.857	0.857
Hindi-BERT	82.73	0.5374	0.680	0.720	0.700
MuRIL + LSTM	86.16	0.4528	0.8266	0.8616	0.8282
Hindi-BERT + LSTM	93.75	0.2341	0.8902	0.8757	0.9106

5.2.5 Sensitivity Analysis and Ablation Study

To assess robustness, we performed a sensitivity analysis on the best-performing model (Hindi-BERT+LSTM). Hyperparameters were varied systematically:

1. ***Learning rate:*** {1e-5, 1e-4 (default), 1e-3}. Validation accuracy at 1e-5 was 87.2% (slower convergence), at 1e-4 was 92.3%, and at 1e-3 was 84.5% (instability). The chosen 1e-4 provided the best trade-off.
2. ***Batch size:*** {16, 32 (default), 64}. Accuracy: 91.1% (batch 16), 92.3% (batch 32), 91.8% (batch 64). Batch 32 yielded the highest accuracy with excellent memory efficiency. - ***Sequence length:*** {128, 256 (default), 512}. Accuracy: 90.5% (128 – information loss), 92.3% (256), 92.1% (512 – no improvement but higher memory). 256 was optimal.
3. ***Ablation study*****:** To quantitatively assess the contribution of the LSTM component, we compared the full Hindi-BERT+LSTM model with two ablated variants:
   i) Hindi-BERT (without using LSTM; only [CLS] token for classification).
   ii) LSTM without BERT (using only static word embeddings from FastText, with same LSTM layers).
4. ***Results:*** Only Hindi-BERT model gave an accuracy of 82.73% (Table 4). The accuracy of LSTM was 71.4%. The complete model (Hindi-BERT+LSTM) performed at 93.75% accuracy which is a relative improvement of 13.3% over Hindi-BERT. This indicates that the LSTM layers provide significant contributions, as they can capture sequential patterns that the transformer's attention mechanism might fail to capture.

5.2.6 Deployment Considerations

The Hindi-BERT+LSTM model requires approximately 500 MB of GPU memory and infers a 256 token tweet in 0.21 seconds on a single T4 GPU (~285 tweets per minute). For larger platforms, model distillation or 8-bit quantization can reduce memory to ~125 MB and increase throughput to ~800 tweets per minute with an accuracy drop of 1–2%. The model can be deployed as a prefilter, automatically flagging high confidence posts for action and routing uncertain cases to human review.

5.3 Error Analysis

To understand the limitations of our best-performing model (Hindi-BERT+LSTM), we manually reviewed 100 misclassified posts from the validation set (approximately 2.5% of validation data). Three common failure modes were identified:

1. Sarcastic hostility (48% of errors): The model failed to detect hate when it was expressed sarcastically. For example, a post saying "बहुत अच्छे काम किए हैं आपने" ("You have done very good work") – in context, the comment was directed at a politician after a scandal and was intended as sarcastic hate. The model classified it as non-hostile because it lacked explicit profanity or hate keywords.
2. Neologisms and slang (32% of errors): Newly emerging derogatory terms not present in the training data (e.g., urban slang combining Hindi and English stems) were missed. For instance, "चपड़ी" (a recent slang for an annoying person) was not recognized as offensive.
3. Long posts with context truncation (20% of errors): Posts longer than 256 tokens (the maximum sequence length) were truncated, losing crucial context. In one example, a 310-token post built a hostile argument gradually, with the hateful conclusion truncated. The model only saw the neutral beginning.

These error patterns suggest that future improvements should focus on: (a) incorporating sarcasm detection modules (e.g., using sentiment shift detection), (b) regularly updating the token vocabulary with emerging slang, and (c) testing longer sequence lengths (e.g., 512 tokens) or hierarchical models.

5.4 Implications for Digital Communication and Language Studies

The findings of this research are of significant value to digital communication, language analysis, and media research, particularly regarding Hindi-language online discourse. By demonstrating the effectiveness of transformer-based models combined with sequential learning techniques, this study provides communication scholars with scalable tools to identify patterns of hostility that shape the tone and quality of digital communication.

For communication researchers, automated detection models enable analysis of hostility levels across large volumes of Hindi social media content – something previously impractical with manual methods. Researchers can now study: (a) how hate speech spreads through network structures; (b) whether certain times of day or events trigger spikes in hostile language; and (c) how linguistic forms of hate differ across platforms (Twitter vs. Instagram vs. Facebook). These capabilities provide new opportunities for research on polarization, digital aggression, and the effect of language on public opinion formation (Papacharissi, 2016; Mathew et al., 2019).

For content moderators and platform designers, the proposed Hindi-BERT+LSTM model offers a practical tool: it achieves 93.75% accuracy on held-out data, meaning that out of 100 hostile posts, approximately 94 would be correctly flagged, with only six false negatives. At the same time, the precision of 0.89 means that when the model flags a post as hostile, it is correct 89% of the time – reducing the burden on human moderators. Platforms can deploy this model as a pre-filter, triaging the most certain hate speech cases for automatic action (e.g., flagging, reduced reach, or removal) while sending uncertain cases for human review.

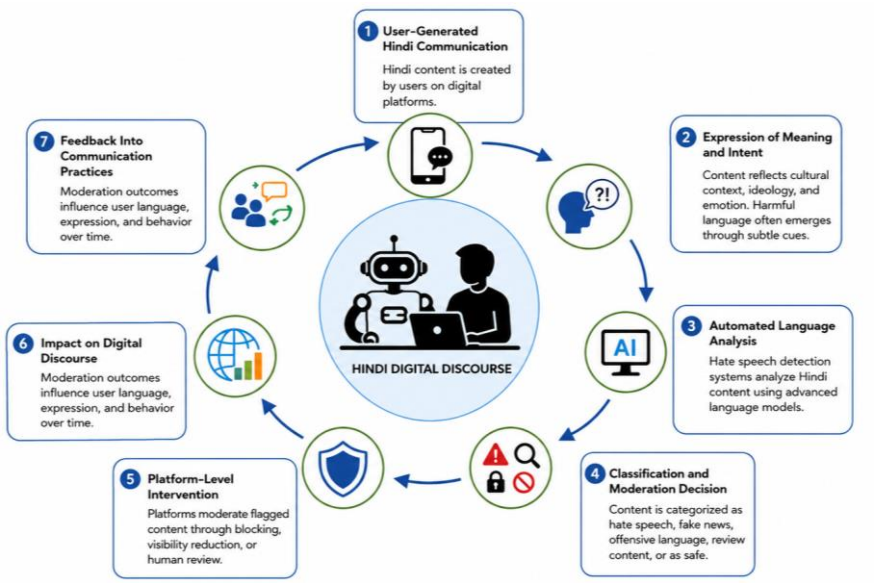
For linguists and cultural analysts, language-specific models like Hindi-BERT are valuable because they respect Hindi's morphology, particle usage, and honorifics – features that general multilingual

models often mishandle. The model's success indicates that computational approaches can detect culturally embedded hostility that keyword-based filters might overlook. For example, sarcastic or indirect hostility (a common failure mode noted in Section 5.3) often relies on cultural knowledge that these models partially learn from training data.

Finally, this study demonstrates how computational language models can augment communication and cultural studies research by providing scalable, replicable methods to analyse complex language phenomena. By combining technical approaches with communication-centred perspectives, researchers can develop a more comprehensive understanding of digital discourse and its societal impacts. Figure 9 depicts the conceptual framework of Hindi digital discourse moderation as proposed in this study.

Figure 9

The conceptual framework of Hindi digital discourse moderation is proposed in this study



5.4.1 Communication Implications of Hindi Hate Speech Detection

The observed effectiveness of the proposed models should also be interpreted in relation to the characteristics of contemporary digital communication environments. Different platforms encourage distinct interaction patterns through variations in anonymity, content visibility, moderation practices, and recommendation mechanisms. These factors influence not only the prevalence of hostile communication but also the linguistic forms through which hostility is expressed.

These findings further highlight the importance of linguistic and cultural contexts in analysing Hindi online discourse. Hindi social media communication frequently involves code-mixing with English, regional linguistic variation, informal expressions, and culturally embedded references. Such features often alter the meaning and intensity of potentially hostile content, making contextual interpretation an important component of automated detection systems.

Beyond content moderation, the ability to identify hostile communication on a large scale provides opportunities to examine broader patterns of online interaction. A large-scale analysis of hostile language may help clarify processes associated with social polarization, formation of opinion clusters, and changes in the quality of public discussion within digital communities.

5.4.2 Methodological Contribution to Communication Research

From a methodological perspective, this study demonstrates how computational techniques can support communication research involving large volumes of user-generated content. Rather than replacing qualitative approaches, such methods can complement traditional discourse analysis by identifying patterns that warrant deeper contextual investigation.

5.5 Ethical Considerations and Bias

Pre-trained language models may encode societal biases present in their training data (Bender et al., 2021; Caliskan et al., 2017). For hate speech detection, this raises two concerns: (a) models might disproportionately flag posts from certain communities as hate speech, and (b) models might miss hate speech directed at communities if such speech was underrepresented in training.

This dataset was annotated by native speakers from three different regions of North India (Delhi, Lucknow, Bhopal). While the approach adds regional diversity, the annotators were drawn from three North Indian cities and brought different regional perspectives. Consequently, the dataset may under-represent perspectives of religious minorities or lower-caste groups regarding what constitutes offensive or hateful speech. Future work should involve annotators from diverse religious, caste, and regional backgrounds.

Additionally, the tweet selection method (keyword and hashtag-based scraping) may introduce sample bias. Tweets that contain hostile keywords are over-represented, while subtly hostile tweets without those keywords are under-represented. This means the model's performance may not generalise to all Hindi social media content, particularly to posts that use indirect or coded hostility.

To mitigate these issues in future deployments, it is recommended to: (a) audit the model's performance across different demographic groups before deployment; (b) use ensemble methods with multiple annotator perspectives; and (c) regularly retrain with updated data to capture evolving language. These steps align with emerging best practices for fair and responsible hate speech detection (Mitchell et al., 2019; Raji et al., 2020).

6.0 Conclusion

Integrating LSTM layers with transformer-based models (Hindi-BERT+LSTM) improved hate speech detection accuracy from 82.73% (Hindi-BERT alone) to 93.75%, confirming the value of sequential learning for Hindi social media text. This hybrid approach also achieved an F1 score of 0.9106, substantially outperforming individual BERT or LLaMA models. For social media platforms, this means more accurate flagging of hostile content with fewer false positives, enabling faster moderation and safer online communities.

One limitation is the risk of overfitting when working with smaller datasets. Additionally, our scope is limited to Hindi content, which may restrict generalisation to other Indian languages or code-mixed inputs. Future work will expand the dataset to include more diverse languages, real world data from multiple platforms, and incorporate sarcasm detection to address the most common failure mode identified in our error analysis.

Future research may extend this work by incorporating sentiment trajectories, social network analysis, and ethnographic approaches to better understand how hostility develops, spreads, and is negotiated within Hindi digital publics. Such interdisciplinary approaches would provide a more holistic understanding of the relationship between language, culture, and digital communication.

Conflict of Interest

The author (s) have declared that no competing interests exist.

Author Contribution Statement

RN and VG: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation. JK and TR: Project Administration, Writing – Review & Editing; VG and VK: Project Administration, Supervision, Writing – Review & Editing.

Funding

This research received no external funding.

Ethics Statement

This research did not require IRB approval because it used publicly available secondary data with no identifying personal information.

Data availability:

The partial data generated during the current study are available from the corresponding author upon reasonable request. The “Hostility Detection Dataset in Hindi” is a publicly accessible dataset from which the remaining information is available at: https://doi.org/10.48550/arXiv.2011.03588

Author Biography

Rachna Narula is an Assistant Professor in the Department of Computer Science and Engineering at Bharati Vidyapeeth's College of Engineering, New Delhi, India. Her research interests include artificial intelligence, machine learning, natural language processing, data analytics, and intelligent computing, with a focus on developing AI-driven solutions for real-world applications.

Vedika Gupta is an Associate Professor at Jindal Global Business School, O.P. Jindal Global University, India. Her research interests include social media analytics, natural language processing, multimodal AI, AI-mediated communication, business analytics, and responsible artificial intelligence. She has authored numerous publications in reputed international journals and conferences.

Jawad Khan is an Assistant Professor with the School of Computing, Gachon University, Republic of Korea. His research focuses on artificial intelligence, deep learning, computer vision, medical image analysis, and intelligent healthcare systems. He has published extensively in high-impact international journals.

Tariq Rahim is an Assistant Professor with the Department of Computer Science, Kingston University London, United Kingdom. His research interests include machine learning, data science, cybersecurity, software engineering, and intelligent information systems, with publications in internationally recognized journals and conferences.

Vijay Kumar is an Assistant Professor in the Department of Computer Science and Engineering at Bharati Vidyapeeth's College of Engineering, New Delhi, India. His research interests include artificial intelligence, data mining, machine learning, and information systems, with contributions to research in intelligent computing and data-driven technologies.

References

Almaliki, M., Almars, A. M., Gad, I., & Atlam, E.-S. (2023). ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media. Electronics, 12(4), 1048. https://doi.org/10.3390/electronics12041048

Alshalan, R., & Al-Khalifa, H. (2020). A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. Applied Sciences, 10(23), 8614. https://doi.org/10.3390/app10238614

Bansal, V., Tyagi, M., Sharma, R., Gupta, V., & Xin, Q. (2022). *A transformer-based approach for abuse detection in code mixed Indic languages*. *ACM Transactions on Asian and Low-Resource Language Information Processing*. Advance online publication. https://doi.org/10.1145/3571818

Bansod, P. P. (2023). *Hate speech detection in Hindi* (Master's project, San José State University). SJSU ScholarWorks. https://doi.org/10.31979/etd.yc74-7qas

Bashar, M. A., & Nayak, R. (2020). *QutNocturnal@HASOC'19: CNN for hate speech and offensive content identification in Hindi language*. *arXiv*. https://doi.org/10.48550/arXiv.2008.12448

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bhardwaj, M., Akhtar, M. S., Ekbal, A., Das, A., & Chakraborty, T. (2020). *Hostility detection dataset in Hindi*. arXiv. https://doi.org/10.48550/arXiv.2011.03588

Bilewicz, M., & Soral, W. (2020). *Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization*. *Political Psychology, 41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). *A dataset of Hindi-English code-mixed social media text for hate speech detection*. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (pp. 36–41). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-1105

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases*. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic minority over-sampling technique*. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Crawford, K., & Gillespie, T. (2016). *What is a flag for? Social media reporting tools and the vocabulary of complaint*. *New Media & Society, 18*(3), 410–428. https://doi.org/10.1177/1461444814543163

De La Peña Sarracén, G. L. (2021). *Multilingual and multimodal hate speech analysis in Twitter*. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 1109–1110). Association for Computing Machinery. https://doi.org/10.1145/3437963.3441668

de Paula, A. F. M., Bensalem, I., Rosso, P., & Zaghouani, W. (2023). *Transformers and ensemble methods: A solution for hate speech detection in Arabic languages*. *arXiv*. https://doi.org/10.48550/arXiv.2303.09823

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fortuna, P., & Nunes, S. (2018). *A survey on automatic detection of hate speech in text*. *ACM Computing Surveys, 51*(4), 1–30. https://doi.org/10.1145/3232676

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). *Large scale crowdsourcing and characterization of Twitter abusive behavior*. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1), 491–500. https://doi.org/10.1609/ICWSM.V12I1.14991

Garg, H., Jhunthra, S., Goel, R., Sharma, S., Gupta, V., Sharma, R., & Dass, P. (2024). *Understanding user polarisation regarding COVID-19 vaccines through social network analysis*. *Journal of Discrete Mathematical Sciences and Cryptography, 27*(8), 2301–2336. https://doi.org/10.47974/jdmsc-1859

Gillespie, T. (2019). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gupta, V., Jain, N., Shubham, S., Madan, A., Chaudhary, A., & Xin, Q. (2021). *Toward integrated CNN-based sentiment analysis of tweets for scarce-resource language—Hindi*. *ACM Transactions on Asian and Low-Resource Language Information Processing, 20*(5), 1–23. https://doi.org/10.1145/3450447

Gupta, S., Nath, T., Gupta, V., & Gupta, M. (2025). *LexiSemIR: A two-stage re-ranking framework with BM25 and zero-shot bi-encoder*. *FIRE 2025 Working Notes*

Hadj Ameur, M. S., & Aliane, H. (2021). *AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset*. *Procedia Computer Science, 189*, 232–241. https://doi.org/10.1016/j.procs.2021.05.086

He, H., & Garcia, E. A. (2009). *Learning from imbalanced data*. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jadhav, I., Kanade, A., Waghmare, V., & Chaudhari, D. (2021). *Hate and offensive speech detection in Hindi Twitter corpus*. In *Working Notes of FIRE 2021: Forum for Information Retrieval Evaluation* (pp. 338–348). CEUR-WS.org. CEUR Workshop Proceedings

Jhunthra, S., Garg, H., & Gupta, V. (2025). *Using tensor processing units to identify the relationship between hypothesis and premise: A case of natural language inference problem*. In S. Dash, B. Naik, & S. K. Shandilya (Eds.), *Uncertainty in computational intelligence-based decision making* (pp. 255–275). Academic Press. https://doi.org/10.1016/B978-0-443-21475-2.00008-4

Kemp, S. (2022, February 15). *Digital 2022: India*. DataReportal. https://datareportal.com/reports/digital-2022-india

Khezzar, R., Moursi, A., & Al Aghbari, Z. (2023). *arHateDetector: Detection of hate speech from standard and dialectal Arabic tweets*. *Discover Internet of Things, 3*(1), Article 18. https://doi.org/10.1007/s43926-023-00030-9

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159. https://doi.org/10.2307/2529310

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). *Spread of hate speech in online social media*. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 173–182). Association for Computing Machinery. https://doi.org/10.1145/3292522.3326034

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting*. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). *A measurement study of hate speech in social media*. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 85–94). Association for Computing Machinery. https://doi.org/10.1145/3078714.3078723

Mossie, Z., & Wang, J.-H. (2018). *Social network hate speech detection for Amharic language*. In D. Nagamalai, M. H. Khafagy, & N. C. S. N. Iyengar (Eds.), *Computer Science & Information Technology (CS & IT): Proceedings of the 6th International Conference on Computer Science and Information Technology (CSIT 2018)* (pp. 41–55). AIRCC Publishing Corporation. https://doi.org/10.5121/csit.2018.80604

Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., & Al-Khalifa, H. (2020, May). Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection* (pp. 48-52)

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., & Abdelali, A. (2021). *Arabic offensive language on Twitter: Analysis and experiments*. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 126–135). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.13/

Nath, T., Gupta, V., Gupta, M., & Sharma, R. (2026). *Decoding multimodal text analytics: Tasks, datasets, fusion models, and future frontiers*. *WIREs Data Mining and Knowledge Discovery, 16*(2), e70083. https://doi.org/10.1002/widm.70083

Nath, T., Singh, V. K., & Gupta, V. (2025). *BongHope: An annotated corpus for Bengali hope speech detection*. *International Journal of Information Technology, 17*(4), 2523–2531. https://doi.org/10.1007/s41870-025-02484-2

Papacharissi, Z. (2016). *Affective publics and structures of storytelling: Sentiment, events and mediality*. *Information, Communication & Society, 19*(3), 307–324. https://doi.org/10.1080/1369118X.2015.1109697

Pawar, A. B., Gawali, P., Gite, M., Jawale, M. A., & William, P. (2022). Challenges for Hate Speech Recognition System: Approach based on Solution. 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 699–704. https://doi.org/10.1109/icscds53736.2022.9760739

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). *Closing the AI accountability gap*. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Raman, S., Gupta, V., Nagrath, P., & Santosh, K. (2022). *Hate and aggression analysis in NLP with explainable AI*. *International Journal of Pattern Recognition and Artificial Intelligence, 36*(15), 2259036. https://doi.org/10.1142/S0218001422590364

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). *A web of hate: Tackling hateful speech in online social spaces*. *arXiv*. https://doi.org/10.48550/arXiv.1709.10159

Schmidt, A., & Wiegand, M. (2017). *A survey on hate speech detection using natural language processing*. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Sharma, D., Gupta, V., & Singh, V. K. (2023). *Detection of homophobia & transphobia in Malayalam and Tamil: Exploring deep learning methods*. In A. Abraham, A. K. Muda, N. Gandhi, S. N. Mohanty, & K. Cengiz (Eds.), *Advanced Network Technologies and Intelligent Computing* (pp. 217–226). Springer. https://doi.org/10.1007/978-3-031-28183-9_15

Sharma, D., Nath, T., Gupta, V., & Singh, V. K. (2025). *Hate speech detection research in South Asian languages: A survey of tasks, datasets and methods*. *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*(3), 1–44. https://doi.org/10.1145/3711710

Sharma, D., Singh, V. K., & Gupta, V. (2024). *TABHATE: A target-based hate speech detection dataset in Hindi*. *Social Network Analysis and Mining, 14*, Article 190. https://doi.org/10.1007/s13278-024-01355-1

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). *Studying generalisability across abusive language detection datasets*. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 940–950). Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1088

Tufekci, Z. (2014). *Social movements and governments in the digital age: Evaluating a complex landscape*. *Journal of International Affairs, 68*(1), 1–18. Journal of International Affairs article

Sahana Udupa. (2018). *Gaali cultures: The politics of abusive exchange on social media*. *New Media & Society, 20*(4), 1506–1522. https://doi.org/10.1177/1461444817698776

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc

Waseem, Z., & Hovy, D. (2016). *Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter*. In *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2013

Zannettou, S., Finkelstein, J., Bradlyn, B., & Blackburn, J. (2020). *A quantitative approach to understanding online antisemitism*. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 786–797. https://doi.org/10.1609/ICWSM.V14I1.7343