

Original Article

GuHPD: A Transformer-Driven Approach to Hostile Post Detection in Gujarati

Jagruiti Boda¹, Keyur Rana²

¹Computer Engineering, Gujarat Technological University, Gujarat, India.

²Computer Engineering, Sarvajani College of Engineering & Technology, Gujarat, India.

¹Corresponding Author : jagrutiboda10@gmail.com

Received: 22 March 2025

Revised: 13 June 2025

Accepted: 16 June 2025

Published: 30 July 2025

Abstract - In recent years, social media has become a prominent medium for public expression; however, it is increasingly exploited to disseminate hostility, particularly against individuals, communities, and religious groups. Religious hate speech can cause profound societal and psychological harm, underscoring the urgent need for automated detection systems. While considerable progress has been made in English-language hate speech detection, limited efforts have addressed low-resource languages such as Gujarati. To bridge this gap, this study presents the Gujarati Hostile Posts Detection (GuHPD) dataset, comprising approximately 14,800 manually annotated comments aimed at identifying hostile content in Gujarati. The dataset supports two core tasks: (i) binary classification to differentiate hostile and non-hostile posts, and (ii) multi-class classification to identify hostile subtypes, including hate speech, fake news, defamation, and offensive language. Annotation reliability was assessed using Fleiss' Kappa, indicating substantial agreement. Several transformer-based models were evaluated, with Multilingual BERT demonstrating the highest performance, achieving an accuracy of 0.93 for binary classification and 0.78 for multi-class classification. These findings demonstrate the utility of the GuHPD dataset in advancing hostile content detection for underrepresented languages and provide a benchmark for future research in regional NLP applications.

Keywords - Coarse-grained text classification, Deep Learning, Fine-grained text classification, Gujarati dataset, Hostile post.

1. Introduction

Social media platforms, now engaging over 4.8 billion users globally [1], have significantly influenced the way individuals communicate, share opinions, and participate in public discourse. However, this rapid digital expansion has also facilitated the widespread circulation of harmful content, including hate speech, misinformation, and targeted hostility. These platforms, including X (formerly Twitter) and YouTube, have increasingly been exploited to promote divisive narratives and attack specific individuals, communities, or religious groups [3]. Such online hostility has been associated with severe societal outcomes. A notable example is the amplification of violence against the Rohingya Muslim minority in Myanmar, where coordinated online hate contributed to real-world atrocities [4]. Incidents like these demonstrate the urgent need for robust automated systems capable of detecting and mitigating hostile content in real-time [5]. While substantial progress has been achieved in detecting hate speech in widely spoken languages such as English [6, 7] and Hindi [8-15], much of this research does not extend to regional or low-resource languages. Code-mixed datasets have also been explored to some extent [17-24], but regional Indian languages like Marathi [25-32], Gujarati [33-35, 39, 42], Telugu [36], Tamil [37], Assamese [38], Sinhala [39],

Bengali [40], and Arabic [41] remain comparatively underrepresented in this domain. The increasing prevalence of regional language content on social media introduces additional challenges in automated hostile post detection. These languages often lack comprehensive annotated datasets and pretrained models. For Gujarati in particular, existing tools are insufficient for the complex linguistic and cultural nuances present in online hostility. This gap highlights the pressing need to develop robust, language-specific detection models and datasets that can address emerging risks in low-resource language contexts.

1.1. Motivation and Research Gap

The proliferation of hostility and hate speech in regional languages on social media, especially Gujarati, has created significant safety concerns. Although more than 62 million people speak Gujarati around the world [43], it is still not well-studied in natural language processing (NLP). Many Gujarati users on social media face harmful content such as hate speech, offensive language, wrong information, and abusive language [3]. However, very little research is available to detect such content in the Gujarati language. Most earlier research has focused on well-known or mixed languages [6-24], often using small datasets and basic classification tasks.



In the case of Gujarati, our past studies usually deal with only two categories (like hostile or not-hostile) using simple machine learning models such as SVM, KNN, RF, etc. [33-35], without exploring detailed types of hostile content.

Also, there is a lack of public datasets for Gujarati, and the ones that do exist often have poor annotated data. This study tries to solve these problems by creating a well-annotated, high-quality dataset for detecting hostile content in Gujarati. It also supports two- and multi-class classification and shares performance results using modern transformer-based models.

1.2. Challenges

There are many challenges in detecting hostile content in the Gujarati language. Some of the key challenges include:

- Low-resource constraints: The language lacks large-scale annotated datasets and pretrained language models specific to its grammar and semantics [44].
- Context sensitivity: Hostile text often relies on cultural context, sarcasm, or complex language, which is difficult to detect using models built for high-resource languages.
- Multilabel complexity: Moving from binary to multi-class classification-such as identifying hate speech, fake news, offensive language, and defamation-brings significant challenges due to overlapping categories and mixed forms of hostility [5].

1.3. Contributions

This research gives the following important contributions:

- Dataset Creation: We developed the Gujarati Hostile Post Detection (GuHPD) dataset by collecting 14,800 comments from X (formerly Twitter). The dataset underwent thorough preprocessing, manual annotation, and evaluation of inter-annotator agreement to ensure high quality. Annotations were done for both binary classification (hostile vs. non-hostile) and multilabel classification, covering categories such as Hate, Fake, Offensive, and Defamation.
- Classification Tasks:
 - Level 1: Binary Classification (Coarse-grained): Differentiating between hostile and non-hostile content.
 - Level 2: Multi-class Classification (Fine-grained): Classifying hostile posts into four specific categories-hate, fake, offensive, and defamation.
- Annotation Agreement calculation: Measured using Fleiss' Kappa to ensure inter-annotator reliability for both binary and multi-class labels.
- Model Benchmarking: Evaluation of various transformer-based models, with Multilingual BERT achieving 93% accuracy on binary classification and 78% on multi-class classification.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes dataset construction and annotation methodology; Section 4 outlines the models used; Section 5 presents the experimental results; Section 6 presents the Result Analysis; Section 7 discusses findings; Section 8 concludes the paper.

2. Related Work

2.1. Hostile Speech in Low-Resource Language

Detecting hostile speech in low-resource languages presents several challenges, such as the use of slang, non-standard spelling, grammatical inconsistencies, and context-dependent expressions like sarcasm or emotional cues. These aspects make it difficult for traditional or generalized models to detect hostile intent accurately. The lack of annotated data and linguistic tools further limits research progress for many Indian regional languages, including Gujarati.

2.2. Existing Research across Languages

In high-resource languages such as Hindi and English, numerous studies have demonstrated the effectiveness of deep learning and transformer-based approaches in classifying hostile content. For Hindi, models like CNN, LSTM, and BERT, combined with embeddings such as FastText and word2vec, have shown considerable success. Several benchmark datasets-like Constraint@AAAI2021-have enabled both coarse-grained and fine-grained hostile post classification tasks [9]. For example, Sreelakshmi K. [45] used FastText along with word2vec and doc2vec embeddings in Hindi-English code-mixed text, finding character-level features more effective than word-level representations.

Kamble [22] applied CNN, LSTM, and BiLSTM for hate speech detection in code-mixed datasets, demonstrating the capability of deep neural models to learn semantic and contextual patterns. Chavan et al. [25] used transformer models such as MuRIL and MahaTweetBERT for offensive language detection in Marathi, achieving high macro F1-scores on datasets like HASOC 2021 and HASOC 2022. Table 1 compares various studies across languages, including their methodology, feature extraction techniques, datasets, evaluation metrics, and target labels. These studies confirm that transformer-based models tend to outperform traditional machine learning techniques, particularly in tasks involving nuanced or context-sensitive content.

2.3. Research Gap

Although Gujarati is India's sixth most spoken language, it remains a low-resource language in terms of publicly available datasets and NLP tools for hostile content detection. Research focusing on Gujarati is limited, with very few contributions addressing even binary classification of hostile content. Our previous study represents one such attempt, which evaluated machine learning models like SVM using Bag-of-Words and TF-IDF for baseline hostile post detection on a 10,000-comment dataset.

Table 1. Comparison of various hostile post detection in various languages

Reference	Language	Approach	Feature Extraction	Dataset	Evaluation Metric	Labels
[45]	Hindi-English (code-mixed)	SVM, RBF	word2vec, doc2vec	Twitter, Facebook (10,000)	Accuracy, F1-Score	Hate, Non-Hate
[25]	Marathi	BERT (MuRIL, MahaTweetBERT)	N/A	HASOC 2021, 2022 (54,970)	Macro F1-Score	Offensive, Non-Offensive
[9]	Hindi	CNN, LSTM, BERT	FastText	Constraint@AAAI2021 (8192)	Weighted F1-Score	Hostile, Non-Hostile, Hate, Defamation, Offensive
[22]	Hindi-English (code-mixed)	CNN-1D, LSTM, BiLSTM	N/A	Twitter (255,309)	F1-Score	Hate, Non-Hate
[41]	Arabic	CNN, RNN, GRU, BERT	Character n-gram	Twitter (9316)	AUROC, F1-Score	Hateful, Abusive
[46]	English	BERT, Roberta, DistilBERT	TF-IDF	COVID-19 datasets	F1-Score	Extremist, Non-Extremist
[40]	Bengali	SVC, RF, CNN-LSTM	TF-IDF	Facebook (42,036)	Accuracy, F1-Score	Political, Religious, Sexual
[10]	Hindi	SVM, KNN, Naive Bayes	BoW, TF-IDF	Collected (5884)	Accuracy, F1-Score	Hate, Non-Hate

That work explored emoji-based analysis and coarse-grained categorization but lacked multi-class classification and advanced model architectures. There is no prior work focusing on fine-grained classification of hostility (e.g., hate, fake, offensive, defamation) in Gujarati. Furthermore, religious hostility, a particularly sensitive category of online hate, has not been examined in the Gujarati context, unlike in other languages where Islamophobia or religious hate is sometimes the sole focus. Existing models and datasets are generally trained for high-resource languages or code-mixed datasets, leaving a noticeable void in Gujarati NLP research.

2.4. Significance of Our Study

To address these gaps, our current work introduces a novel annotated dataset designed specifically for hostile post detection in Gujarati. This dataset includes approximately 14,800 comments manually labeled for both binary and fine-grained multi-class classification. It is the first of its kind to focus on hostile content targeting religious groups in the Gujarati language, thus opening new avenues for research and practical deployment of content moderation tools for this linguistic community. By evaluating transformer-based models on this dataset and benchmarking their performance, we provide essential resources and insights that can guide future developments in hate speech detection for underrepresented languages.

3. Dataset Curation

3.1. Data Collection and Pre-Processing

We collected X (formerly known as Twitter) data using 90 distinct keywords chosen to capture significant events in Gujarat over the past five years—Figures 1 and 2 present word

clouds of these keywords in English and Gujarati, respectively. To ensure balanced data, we collected an equal number of comments for each keyword, following X's limit of 3,200 comments per keyword search. The data collection spanned from 2017 to 2022 using these identified keywords.

The comments were then compiled into a single CSV file for model training and analysis. The dataset captures user comments on both positive and negative events in Gujarat over the past five years, covering various domains such as Sports and Entertainment, Government and Politics, International Relations, Literature and Safety, Environment and Climate, Law and Justice, Economy and Business, and Miscellaneous Events, as depicted in Figure 3.

The preprocessing is necessary due to the high noise in the data from X (formerly known as Twitter). The model preprocessing steps are explained in detail in our previous research. Data augmentation through translation is employed to increase the training dataset and improve classification performance. Posts are first translated from Gujarati to English and then back to Gujarati using the Google Translate API [48], creating diverse linguistic expressions while maintaining the original context.

To evaluate the accuracy of the translation, the Levenshtein distance [47] is used to measure the similarity between the original and back-translated Gujarati posts. The average translation accuracy score, based on the Levenshtein distance, is 74.32%. This score reflects the reliability of the augmented data in preserving the core meaning, which in turn aids in enhancing the classification model's performance.



Fig. 1 Wordcloud of Gujarati good and bad keywords in Gujarati

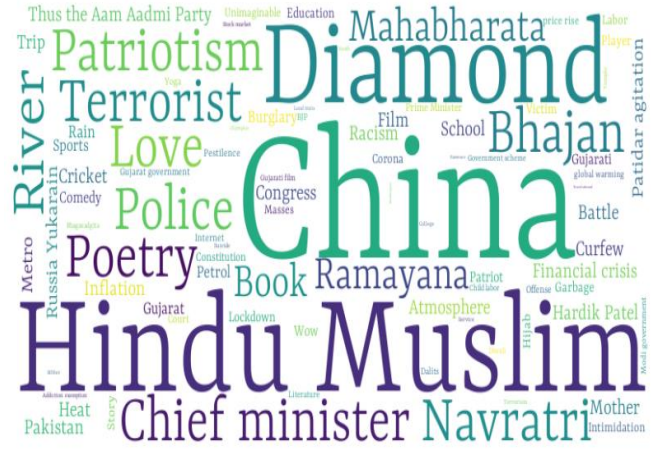


Fig. 2 Wordcloud of Gujarati good and bad keywords in English

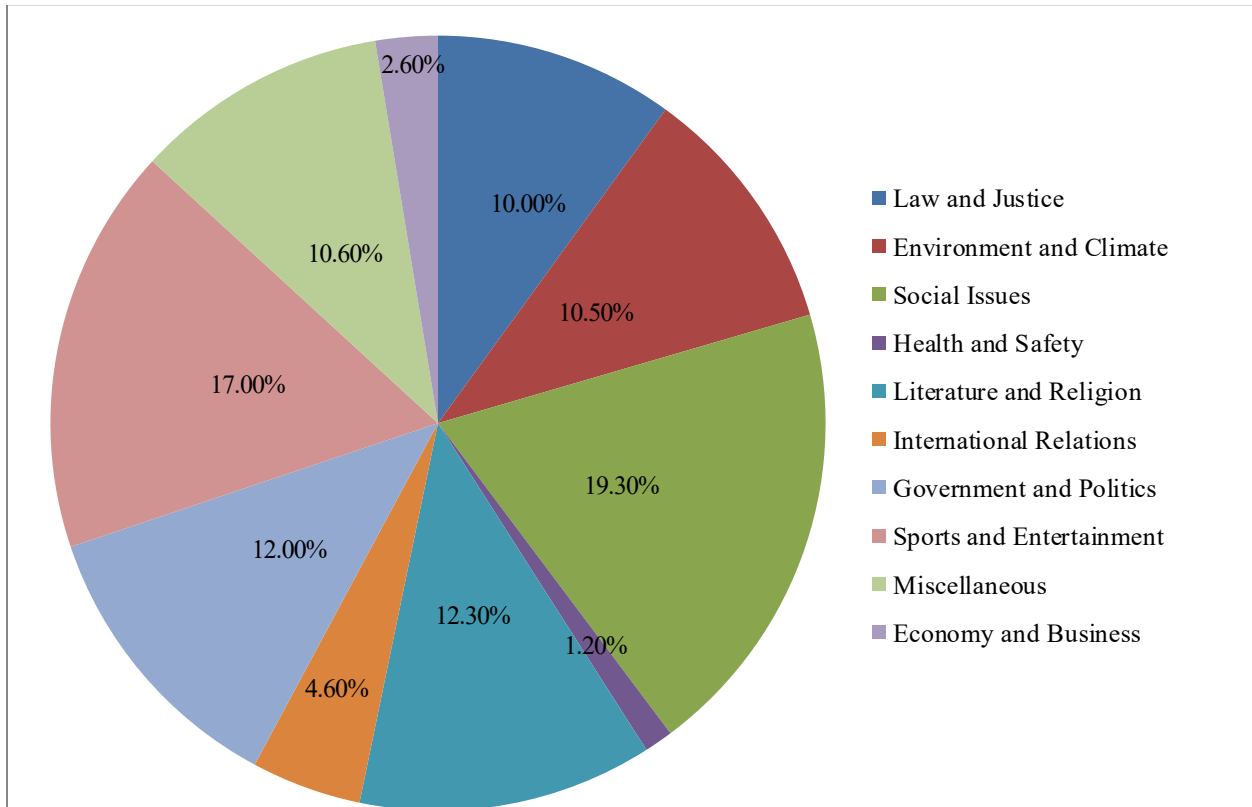


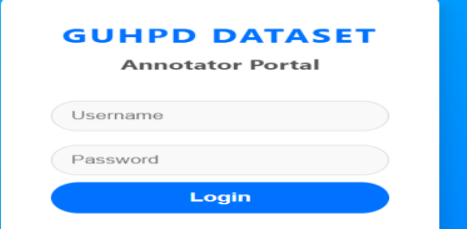
Fig. 3 Data collection domain

3.2. Data Annotation

Figure 4 displays the login page of the application, where users enter their credentials to access the system.

Figure 5 shows the data annotation page for the Non-Hostile category, where users annotate comments as non-hostile.

Figure 6 presents the data annotation page for the Hostile category, where users classify comments as a hostile subcategory (Hate, Fake, Offensive and Defamation).



GUHPD DATASET

Annotator Portal

Username

Password

Login

Fig. 4 Login page

ANNOTATE COMMENT

નવસારી ખાતે આમ આદમી પાર્ટી દ્વારા પરિવર્તન યાત્રાનું આયોજન

Subtask 1: Select Category

Non-Hostile

Submit Annotation

Fig. 5 A screenshot of the data annotation page (non-hostile category)

ANNOTATE COMMENT

BJP સરકાર માં પોલીસ ને પોતાની માં બહેનો ની પણ ઇજ્જત લુંટવાની સત્તા છે. જયભારત

Subtask 1: Select Category

Hostile

Subtask 2: Select Hostile Type

Offensive

Submit Annotation

Fig. 6 A screenshot of the data annotation page (hostile category)

The following annotation schema was adopted for our research, guiding annotators through two subtasks to classify comments into appropriate categories:

3.2.1. Subtask 1 (Binary Classification)

For the first subtask, comments are classified into two categories:

Hostile

Comments fall under the 'Hostile' category if they meet one or more of the following criteria:

- Comments that exhibit aggressive or provocative behavior intended to incite conflict.
- Comments that aim to incite violence or hostility towards individuals or groups based on identity attributes.
- Comments that use disrespectful, vulgar, or inflammatory language aimed at degrading individuals or groups.

Non-Hostile

Comments that do not show hostility or aggression, are respectful in tone, and do not provoke any form of hatred or violence.

3.2.2. Subtask 2 (Multi-class Classification)

If a comment is labeled as 'Hostile' in Subtask 1, it is further classified into one or more of the following categories:

Hate Post

Comments that show anger, encourage violence, or create dislike against certain groups based on their religion, caste, race, or area. These posts may also try to cause trouble or increase conflict in society.

Fake Post

Statements or claims that are not true and do not have real facts, no matter what the speaker means. These often cause false information to spread and confuse people.

Offensive Post

Comments that have rude, insulting, or bad language aimed at people or groups. These remarks may try to shame, make fun of, or disrespect others without always causing hate.

Defamation Post

Comments that include false or unsupported claims about a person or organization meant to harm their reputation or trustworthiness.

Notes for Annotators,

- Some comments may be assigned to multiple subcategories. For instance, a single post may include both offensive and hateful elements.
- Comments identified as *Non-hostile* in Subtask 1 should not proceed to this stage, as they do not exhibit harmful or aggressive intent.

This classification scheme ensures a systematic and detailed annotation process, enabling more accurate modeling of various forms of online hostility in the Gujarati language context.

3.3. Example of Annotated Dataset

This section explains examples from both subtasks to enhance comprehension.

3.3.1. Subtask 1

Hostile

Comment: “BJP સરકાર માં પોલીસ ને પોતાની માં બહેનો ની પણ ઇજ્જત લુંટવાની સત્તા છે. જયભારત”

Translation: "In the BJP government, the police have the power to disrespect their mothers and sisters. Jai Bharat."

Category: Hostile.

Explanation: This statement demonstrates aggressive and accusatory language directed at authorities, classifying it as hostile.

Non-Hostile

Comment: “આમ આદમી પાર્ટી મોરબી દ્વારા પ્રભાબેન રમેશભાઈને ઉપપ્રમુખની જવાબદારી સોંપાઈ”

Translation: "The Aam Aadmi Party has assigned Prabhabeen Rameshbhai the responsibility of the Vice President in Morbi."

Category: Non-hostile

Explanation: This comment simply states a fact about a political appointment without.

Hate

Comment: "જમ્મુ કાશ્મીર: અનંતનાગ એન્કાઉન્ટરમાં 2 આતંકવાદી ઠાર મરાયા, સર્ચ ઓપરેશન હજુ પણ શરૂ #Jammukashmir #CGNews"

Translation: "Jammu Kashmir: 2 terrorists were killed in an encounter in Anantnag, search operation still ongoing."

Category : Hate

Explanation: This comment mentions violence and terrorism, reflecting an incitement of hate towards a particular group.

Fake

Comment: "ફ્રી...ફ્રી...ફ્રી...ફ્રી...ફ્રી... આમ આદમી પાર્ટી તરફથી ગુજરાતની જનતા માટે ખાસ ફ્રી... ફ્રી...ફ્રી...કોમ્પો ઓફર... આમ આદમી પાર્ટી ને વોટ આપો અને મફત ૩૦૦ યુનિટ વીજળી જોડે વિદ્યુત્તી નો આતંક ફ્રી..."

Translation: "Free...Free...Free...Free...Free... A special free offer from the Aam Aadmi Party for the people of Gujarat.

Vote for the Aam Aadmi Party and get 300 units of free electricity along with free terror from non-Hindus..."

Category: Fake

Explanation: This comment contains exaggerated claims about offers from a political party, presenting misinformation.

Offensive

Comment: "@BJPHardikPatell આમ આદમી પાર્ટી ઠગ પાર્ટી છે..🙄"

Translation: "@BJPHardikPatell The Aam Aadmi Party is a cheating party..🙄"

Category: Offensive

Explanation: This comment uses disrespectful language to describe a political party, expressing disdain.

Defamation

Comment: "@Wish_3025 એના નંબર સરકારી મુતરડી મા લખી નાખવા.. પછી એ જાણે ને એનો ભાવ..... નોટી અમેરિકા.. 🤔🤔🤔🤔🤔🤔🤔🤔"

Translation: "@Wish_3025 Write his number in the government office... then he will know his worth..... dirty American.. 🤔🤔🤔🤔🤔🤔🤔🤔"

Category: Defamation

Explanation: This comment makes derogatory remarks about an individual, harming their reputation.

3.4. Inter-Annotation Agreement

Before we use the GuHPD dataset for research on hostile posts, it is important to check how well the annotations were done. This check helps us understand how consistently the annotators labeled the comments. We use a method called inter-annotator agreement to measure this, which looks at two main things:

1. How much agreement or disagreement there is among the annotators when they label the comments.
2. How much of the agreement or disagreement might happen by chance?

To measure this, we use Fleiss' Kappa [49], a tool for checking agreement when three or more annotators work with different categories. Since our dataset has two tasks with different categories-two in the first task and four in the second-we use Fleiss' Kappa to calculate the level of agreement among the annotators. The Fleiss' Kappa score ranges from -1 to 1, providing a measure of inter-annotator agreement. A score of 0 or lower indicates that there is no agreement among the annotators. Scores between 0.01 and 0.20 reflect slight agreement, while scores ranging from 0.21 to 0.40 suggest fair agreement. Moderate agreement is represented by scores from 0.41 to 0.60, and scores between 0.61 and 0.80 indicate substantial agreement. Finally, scores ranging from 0.81 to 1.00 signify almost perfect agreement among the annotators. In our case, the Fleiss' Kappa score for the first task was 0.69, showing good agreement among the annotators. For the second task, the score was 0.56, reflecting moderate agreement. These results show that while the annotators generally agreed, there was still some difference in how they interpreted the comments. This highlights the need for ongoing training and improvements to the annotation guidelines to make the labeling process even better.

3.5. Dataset Statistics

Table 2 indicates the distribution of comments across different categories, showing the initial dataset sizes, sizes after data augmentation, and sizes after removing duplicates for each category.

The dataset contains a total of 14,682 comments categorized into six distinct types: Non-Hostile, Hostile, Fake, Offensive, Defamation, and Hate. It provides insights into the

prevalence of these categories, along with data augmentation and deduplication statistics to enhance the dataset for analysis.

Table 2. Dataset count category-wise for augmented and non-augmented data

Data category	Dataset size	Dataset size after augmentation	Data size after removing duplicates
Non-Hostile	7400	14800	14749
Hostile	7282	14564	12865
Fake	244	488	487
Offensive	348	696	683
Defamation	511	1022	1012
Hate	964	1928	1913

Figure 7 illustrates the character and word count of posts within the dataset, providing an overview of the length and verbosity of the comments analysed. Figure 8 displays the

count of punctuation marks, hashtags, and mentions present in social media posts, highlighting the usage of these elements in the dataset.

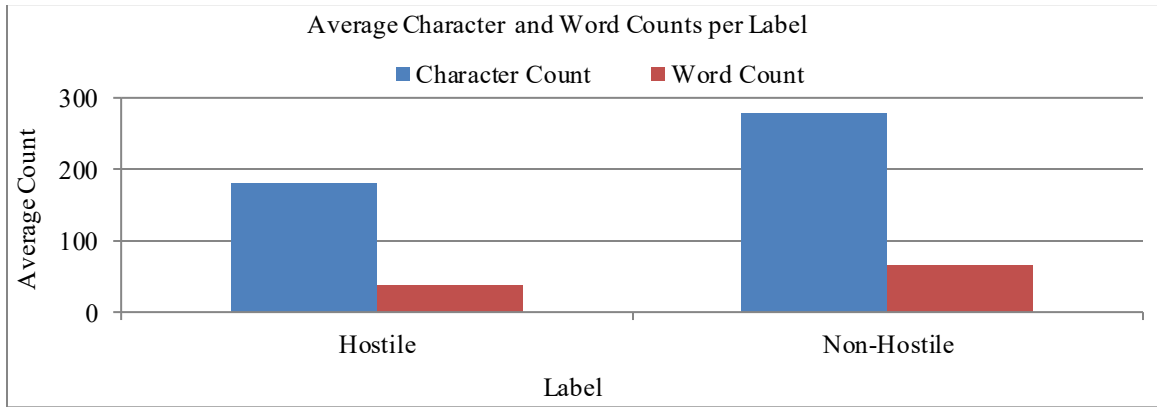


Fig. 7 Character and word count of the post

4. Computational Models

Figure 9 illustrates the block diagram of the proposed approach for hostile post detection. It outlines the key steps, from data preprocessing to classification using transformer-based models for detecting hostile content. We used several state-of-the-art models, including DistilBERT, mBERT,

XLNet, DeBERTa, RoBERTa, and Gujarati-specific models like GujaratiBERT. These models were selected for their robustness in NLP tasks. However, the best-performing model for our data is still unknown. The results of our experiments will determine which model performs best for hostile post detection in Gujarati.

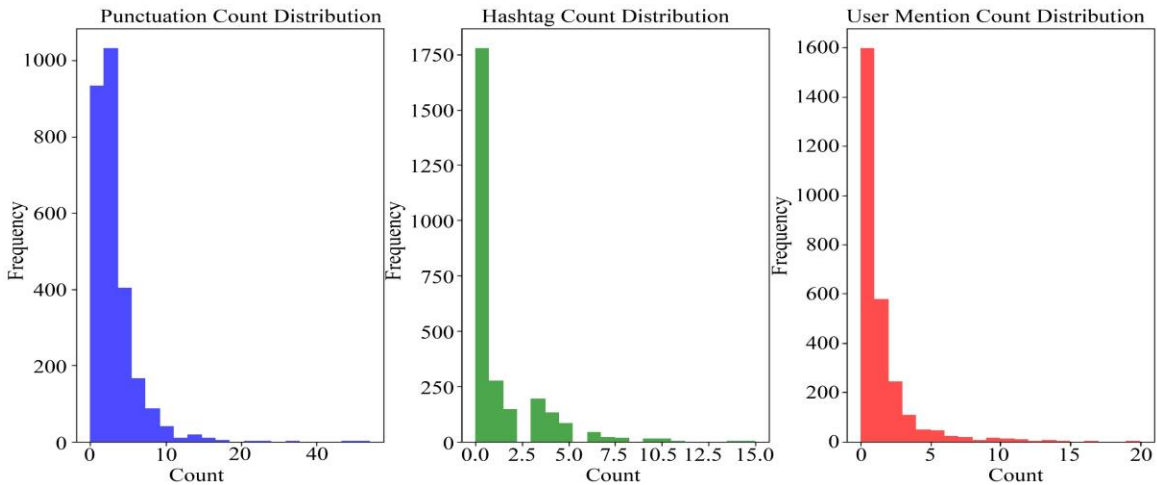


Fig. 8 Count of punctuation, hashtags and mentions in social media post

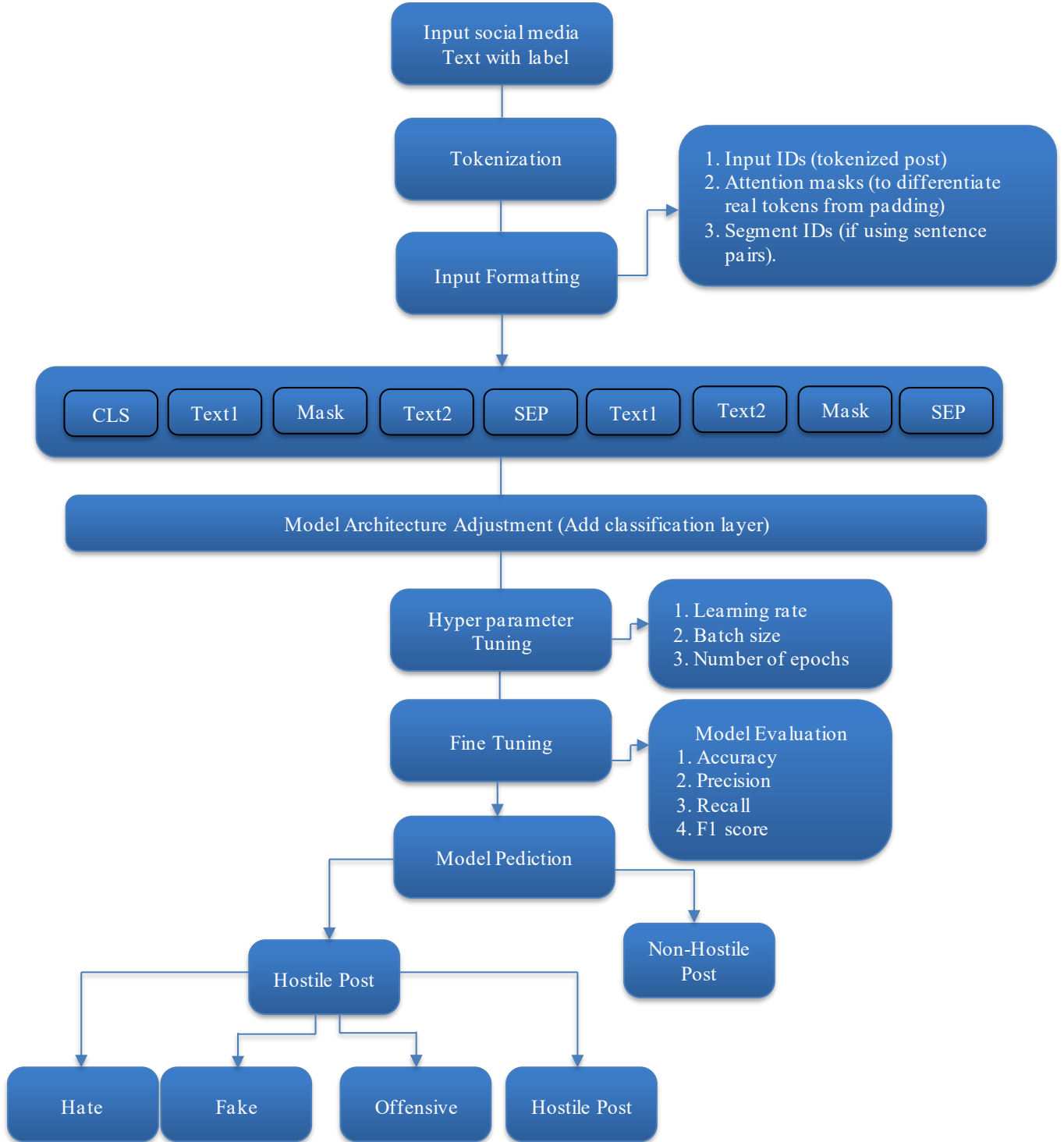


Fig. 9 Block diagram of proposed approach

4.1. DistilBERT [50]

A smaller, faster, and lighter version of BERT, DistilBERT retains 97% of BERT's language understanding while being 60% faster and requiring fewer parameters. It is pre-trained using a knowledge distillation method to compress BERT's architecture while maintaining accuracy.

4.2. MultilingualBERT(mBERT) [51]

mBERT is a multilingual version of BERT pre-trained on the Wikipedia corpus of 104 languages, including Gujarati. It supports text classification and other NLP tasks across multiple languages, providing strong performance in cross-lingual tasks.

4.3. XLNet [52]

XLNet is an extension of the Transformer-XL architecture that incorporates bidirectional context learning. Unlike BERT, which predicts masked tokens independently, XLNet uses a permutation-based training method to capture dependencies between tokens.

4.4. DeBERTa [53]

Decoding-enhanced BERT with disentangled attention (DeBERTa) improves on BERT by using disentangled attention mechanisms and enhanced position encoding, resulting in better performance on various NLP benchmarks compared to BERT and RoBERTa.

4.5. RoBERTa [54]

Robustly optimized BERT approach (RoBERTa) improves BERT by optimizing hyperparameters and training on a larger dataset for longer periods. It removes the Next Sentence Prediction task and uses dynamic masking, achieving better performance on several NLP tasks.

4.6. GujaratiBERT [55]

GujaratiBERT is a BERT model made especially for the Gujarati language. It is trained on a large amount of Gujarati text, which helps it understand the language better than general multilingual models. This model is created to work well on tasks that involve the Gujarati language In Natural Language Processing (NLP).

4.7. L3cube-Pune/Gujarati-Bert [56]

This BERT model is specially fine-tuned for Gujarati by L3Cube Pune. It is trained on a large amount of Gujarati text and is useful for tasks like sentiment analysis and text classification in Gujarati.

4.8. L3cube-Pune/Gujarati-Sentence-Bert-Nli [57]

This version of BERT is fine-tuned for sentence-level tasks in Gujarati, such as understanding sentence meaning (natural language inference) and finding sentence similarities. It uses sentence embeddings to improve performance on these tasks.

4.9. Google-Bert/Bert-Base-Multilingual-Cased [51]

This model is the cased version of Google's multilingual BERT (mBERT), trained on texts from more than 100

languages, including Gujarati. It keeps uppercase and lowercase letters separate and uses data from many languages. This helps it work well on different language tasks and makes it useful for handling multiple languages at once.

4.10. Google/Muril-Base-Cased [59]

MuRIL is a multilingual model trained by Google specifically for Indian languages. It is trained on both translation and transliteration tasks and includes 17 Indian languages along with English, improving on mBERT for Indian language tasks.

4.11. L3cube-Pune/Marathi-Bert [60]

A BERT model fine-tuned for Marathi language tasks, similar to l3cube-pune/gujarati-bert. It is trained on a large Marathi dataset and is designed for various NLP applications like sentiment analysis and text classification in Marathi.

4.12. Google-Bert/Bert-Large-Uncased [61]

This is a larger version of the original BERT model with more layers and parameters (24 layers and 340M parameters), trained on uncased text (where lowercase and uppercase letters are treated the same), making it suitable for more complex tasks requiring deeper language understanding.

5. Results

In Table 3, Multilingual BERT demonstrates the best performance for coarse-grained classification of Gujarati text. It achieves the highest accuracy, precision, recall, and F1 score, reaching an accuracy of 0.89 at Epoch 4, with precision, recall, and F1 scores of 0.86, 0.93, and 0.88, respectively. This outstanding performance indicates that Multilingual BERT effectively handles the linguistic nuances and complexity present in Gujarati hostile post detection. While models like DistilBERT and GujaratiBERT also show strong and stable results, their scores are slightly lower compared to Multilingual BERT. Models such as XLNet and DeBERTa perform relatively worse, suggesting they may not generalize as effectively on this dataset. Figure 10 shows a comparison of how different transformer-based models perform on Gujarati text classification at the coarse-grained level. The chart shows all models' accuracy, precision, recall, and F1 score, making it easy to compare their results. Multilingual BERT gives the best result, showing that it works well for this task.

Table 3. Result of the transformer model for coarse-grained classification on Gujarati data

BERT model	Epoch	Accuracy	Precision	Recall	F1 score
Distilbert [50]	1	0.87	0.88	0.85	0.87
	2	0.88	0.86	0.90	0.88
	3	0.88	0.85	0.92	0.88
	4	0.88	0.84	0.93	0.88
Multilingual [51]	1	0.87	0.85	0.91	0.88

	2	0.88	0.85	0.93	0.88
	3	0.88	0.85	0.92	0.89
	4	0.89	0.86	0.93	0.89
Xlnet [52]	1	0.79	0.74	0.89	0.81
	2	0.78	0.78	0.78	0.78
	3	0.79	0.78	0.82	0.81
	4	0.81	0.79	0.81	0.81
Deberta [53]	1	0.66	0.61	0.92	0.73
	2	0.71	0.66	0.84	0.74
	3	0.69	0.64	0.87	0.74
	4	0.87	0.70	0.88	0.75
Roberta [54]	1	0.72	0.65	0.95	0.77
	2	0.75	0.69	0.89	0.78
	3	0.76	0.70	0.91	0.79
	4	0.77	0.72	0.88	0.79
GujaratiBERT [55]	1	0.86	0.86	0.86	0.86
	2	0.86	0.86	0.86	0.86
	3	0.85	0.85	0.85	0.85
	4	0.86	0.86	0.86	0.86
l3cube-pune/gujarati-bert [56]	1	0.79	0.79	0.80	0.79
	2	0.80	0.80	0.79	0.80
	3	0.80	0.79	0.80	0.80
	4	0.80	0.80	0.80	0.80
l3cube-pune/gujarati-sentence-bert-nli [57]	1	0.76	0.76	0.76	0.76
	2	0.76	0.76	0.76	0.76
	3	0.77	0.77	0.77	0.77
	4	0.77	0.77	0.77	0.77
google-bert/bert-base-multilingual-cased [51]	1	0.66	0.66	0.66	0.66
	2	0.67	0.67	0.67	0.67
	3	0.67	0.67	0.67	0.67
	4	0.67	0.67	0.67	0.67
google/muril-base-cased [59]	1	0.49	0.49	0.49	0.49
	2	0.50	0.50	0.50	0.50
	3	0.49	0.49	0.49	0.49
	4	0.50	0.50	0.50	0.50
l3cube-pune/marathi-bert [60]	1	0.66	0.66	0.66	0.66
	2	0.66	0.66	0.66	0.66
	3	0.67	0.67	0.67	0.67
	4	0.67	0.67	0.67	0.67
google-bert/bert-large-uncased [61]	1	0.50	0.50	0.50	0.50

	2	0.51	0.51	0.51	0.51
	3	0.50	0.50	0.50	0.50
	4	0.51	0.51	0.51	0.51

We perform various experiments to fine-tune the hyperparameters of the Multilingual BERT (mBERT) model. The tuning process involved adjusting the learning rate, batch size, number of epochs, optimizer, and loss function to improve model performance. Learning rates tested included $1e-5$, $2e-5$, $3e-5$, $4e-5$, and $5e-5$, and we experimented with batch sizes of 32 and 64. The AdamW optimizer was selected because of its ability to manage sparse gradients and apply weight decay effectively, which is beneficial for training transformer models like mBERT. For the loss function, we used cross-entropy loss, as it is well-suited for classification tasks and helps in evaluating how well the predicted class probabilities match the actual labels. Other optimizers, such as

Stochastic Gradient Descent (SGD) [2] and Root Mean Square Propagation (RMSprop) [58], were not chosen due to slower convergence rates. Additionally, mean squared error [16] was deemed unsuitable for classification tasks, as it is more appropriate for regression problems. Additionally, we explored the impact of data augmentation to improve generalization. We conducted experiments for both coarse-grained binary classification (Hostile vs. Non-hostile) and fine-grained multilabel classification (Hate, Fake, Offensive, and Defamation). The experiments provide insights into how these factors affect mBERT's performance across multilingual tasks. Detailed results and findings are presented in the following sections.

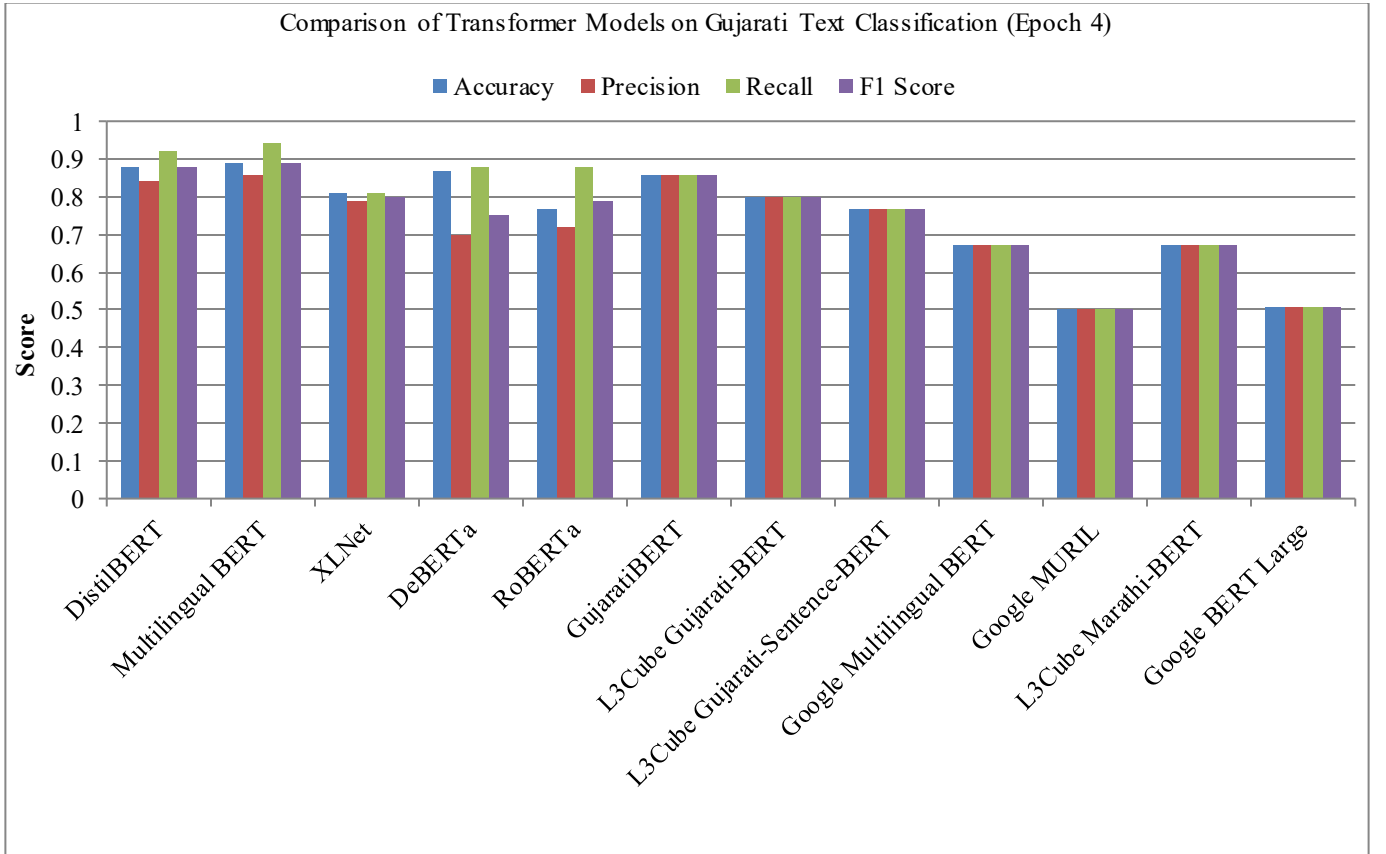


Fig. 10 Comparison of transformer models on Gujarati hostile post detection

5.1. Experiment for Coarse-Grained Classification

5.1.1. Experimental Result 1: Analyze the Effect of Learning Rate on Model Fine-Tuning

We performed experiments to determine the optimal learning rate for fine-tuning a pretrained BERT model on the hostile post detection task, testing values between $1e-5$ and $5e-5$. Table 4 summarizes the model's performance across these

learning rates. The learning rate of $2e-5$ consistently yielded the highest accuracy, precision, recall, and F1 score across multiple epochs, showing steady improvement with minimal training loss. In contrast, higher learning rates, such as $3e-5$ and $5e-5$, caused slight fluctuations in performance. Thus, $2e-5$ proved to be the most effective, offering the best balance between model performance and stability.

Table 4. Experiment result for coarse-grained classification with changing learning rate

Epoch	Learning Rate	Accuracy	Precision	Recall	F1 score	Training Lose
1	1e-5	0.87	0.82	0.94	0.88	0.36
2	1e-5	0.87	0.85	0.91	0.88	0.22
3	1e-5	0.88	0.85	0.92	0.88	0.18
4	1e-5	0.88	0.85	0.92	0.88	0.15
1	2e-5	0.86	0.84	0.89	0.87	0.35
2	2e-5	0.87	0.85	0.90	0.88	0.23
3	2e-5	0.88	0.86	0.90	0.88	0.19
4	2e-5	0.88	0.87	0.90	0.88	0.16
1	3e-5	0.87	0.86	0.89	0.88	0.32
2	3e-5	0.87	0.87	0.87	0.87	0.23
3	3e-5	0.87	0.86	0.90	0.88	0.19
4	3e-5	0.88	0.85	0.91	0.88	0.16
1	4e-5	0.86	0.80	0.95	0.87	0.38
2	4e-5	0.87	0.81	0.95	0.88	0.25
3	4e-5	0.87	0.82	0.95	0.88	0.21
4	4e-5	0.88	0.83	0.95	0.88	0.19
1	5e-5	0.87	0.85	0.89	0.87	0.34
2	5e-5	0.87	0.83	0.94	0.88	0.24
3	5e-5	0.88	0.83	0.94	0.88	0.20
4	5e-5	0.87	0.85	0.91	0.88	0.18

Table 5. Model performance check by increasing the epoch

Epoch	Learning Rate	Accuracy	Precision	Recall	F1 score	Training Lose
5	2e-5	0.88	0.87	0.91	0.88	0.13
10	2e-5	0.88	0.86	0.91	0.88	0.04
15	2e-5	0.88	0.86	0.91	0.89	0.02
20	2e-5	0.88	0.87	0.91	0.89	0.017
25	2e-5	0.88	0.87	0.91	0.89	0.015

5.1.2. Experimental Result 2: Analyze the Effect of Epoch on Model Fine-Tuning

Table 5 presents the results of the model's performance with varying epochs. Increasing epochs improves metrics like accuracy, precision, recall, and F1 score, with performance stabilizing around 25 epochs. We chose 25 epochs as it balances training time and model effectiveness. The model does not show signs of overfitting or underfitting, as the training loss decreases consistently while performance metrics remain stable. Therefore, the model fits this epoch value well.

5.1.3. Experimental Result 3: Analyze the Effect of Batch Size on Model Fine-Tuning

Table 6 shows that the model performs similarly with both batch sizes (64 and 32), maintaining high accuracy, precision, recall, F1 score, and ROC AUC score. Batch size 64 consistently shows slightly lower training loss compared to batch size 32. However, the performance metrics remain stable across epochs, indicating that both batch sizes are effective. Therefore, batch size 64 is preferable due to slightly faster convergence with similar performance.

Table 6. Model performance check by changing the batch size value

Epoch	Batch Size	Accuracy	Precision	Recall	F1 score	Training Lose	ROC AUC Score
1	64	0.88	0.84	0.94	0.88	0.37	0.88
2	64	0.88	0.83	0.95	0.89	0.23	0.88
3	64	0.88	0.84	0.94	0.89	0.18	0.88
4	64	0.88	0.84	0.94	0.89	0.15	0.88
1	32	0.86	0.84	0.89	0.87	0.35	0.86
2	32	0.87	0.85	0.90	0.88	0.23	0.87
3	32	0.88	0.86	0.90	0.88	0.19	0.88
4	32	0.88	0.87	0.90	0.88	0.16	0.88

5.1.4. Experimental Result 4: Analyze the Effect of Data Augmentation on Model Fine-Tuning

Table 7 and Figure 11 illustrate the impact of data augmentation on the fine-tuning performance of the model across various epochs. Data augmentation consistently enhances all key evaluation metrics-accuracy, precision, recall, and F1 score-compared to training without augmentation. This improvement is particularly noticeable at higher epochs (15 and 20), where the model trained with augmented data achieves peak performance, with accuracy reaching 0.93 and F1 score at 0.93. The augmentation strategy

effectively mitigates class imbalance by providing the model with more diverse examples of minority classes (hostile posts) often underrepresented in real-world datasets. As a result, the model generalizes better and reduces bias toward majority classes. This is further supported by the decreasing training loss values when data augmentation is applied, indicating improved learning stability and convergence. These results demonstrate that data augmentation plays a vital role in enhancing model robustness and balanced performance, especially in tasks such as hostile post detection, where class distribution is skewed.

Table 7. Model performance comparison for augmented and non-augmented data

Epoch	Data Augmentation	Accuracy	Precision	Recall	F1 score	Training Lose
5	No	0.88	0.87	0.91	0.88	0.13
5	Yes	0.90	0.89	0.92	0.90	0.08
10	No	0.88	0.86	0.91	0.88	0.04
10	Yes	0.92	0.91	0.93	0.92	0.02
15	No	0.88	0.86	0.91	0.89	0.02
15	Yes	0.93	0.92	0.94	0.93	0.01
20	No	0.88	0.87	0.91	0.89	0.01
20	Yes	0.93	0.92	0.94	0.93	0.01

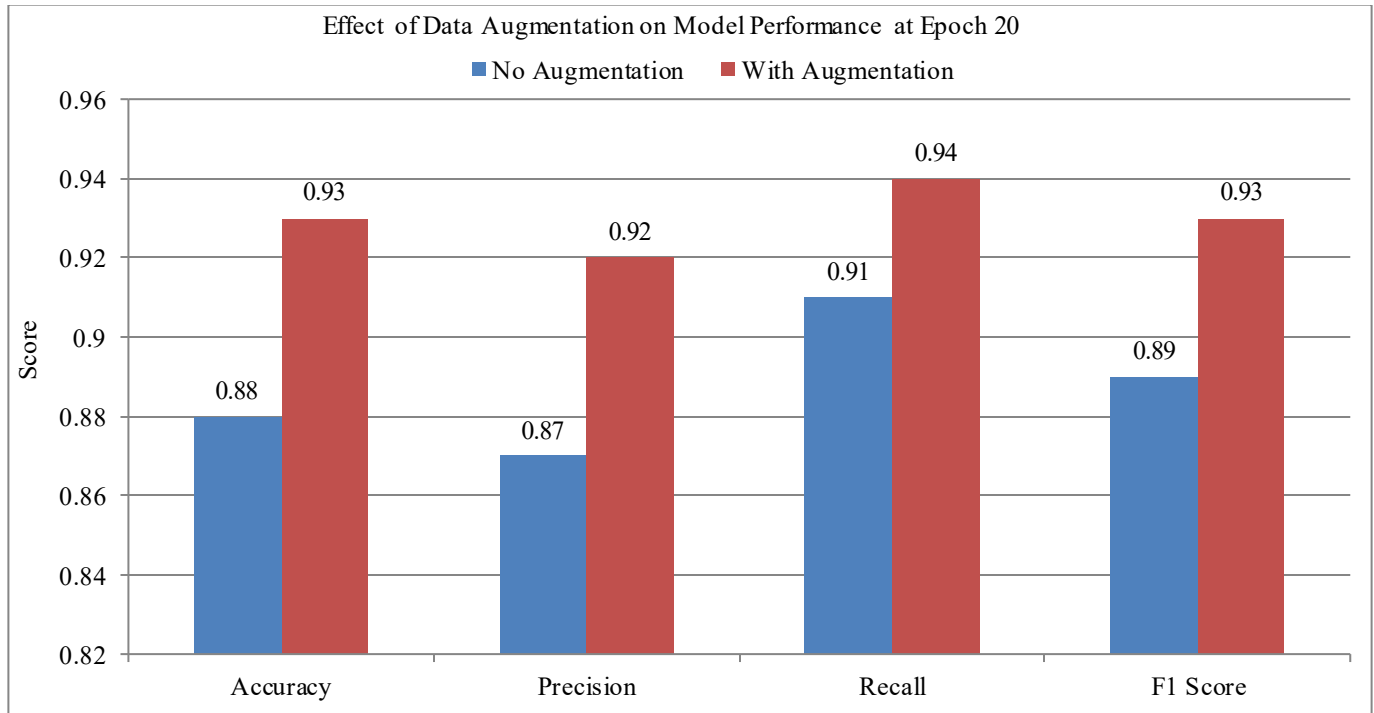


Fig. 11 Impact of data augmentation on model performance at epoch 20, showing improved accuracy, precision, recall, and f1 score

5.2. Experiment for Fine-Grained Classification

In this section, we present the experimental results for fine-grained classification.

5.2.1. Experimental Result 1: Analyze the Effect of Data Augmentation on Model Fine-Tuning

Table 8 and Figure 12 present a comparative analysis of model performance with and without data augmentation over

10 epochs of fine-tuning. The results demonstrate that applying data augmentation significantly enhances the model's learning process and overall performance metrics. Specifically, the model trained with data augmentation shows consistent improvement across all evaluation metrics-accuracy, precision, recall, and F1 score-culminating in a highest accuracy of 0.70 and a training loss of 0.11 at the 10th epoch.

In contrast, the model trained without data augmentation peaks at a maximum accuracy of 0.57 with a noticeably higher training loss of 0.29 at the same epoch. This comparison indicates that data augmentation not only accelerates convergence during training but also improves the model's

ability to generalize to unseen data. The performance gap becomes particularly pronounced in the later epochs, underscoring the effectiveness of data augmentation in fine-tuning deep learning models for fine-grained classification tasks.

Table 8. Model performance for fine-grained classification without augmented data and with augmented data

Epoch	Data Augmentation	Accuracy	Precision	Recall	F1 score	Training Lose
1	No	0.48	0.48	0.48	0.48	1.26
2	No	0.48	0.48	0.48	0.48	1.20
3	No	0.52	0.52	0.52	0.52	1.14
4	No	0.54	0.54	0.54	0.54	0.99
5	No	0.55	0.55	0.55	0.55	0.87
6	No	0.55	0.55	0.55	0.55	0.71
7	No	0.56	0.56	0.56	0.56	0.60
8	No	0.56	0.56	0.56	0.56	0.47
9	No	0.57	0.57	0.57	0.57	0.36
10	No	0.57	0.57	0.57	0.57	0.29
1	Yes	0.55	0.55	0.55	0.55	1.22
2	Yes	0.57	0.57	0.57	0.57	1.02
3	Yes	0.59	0.59	0.59	0.59	0.85
4	Yes	0.61	0.61	0.61	0.61	0.66
5	Yes	0.63	0.63	0.63	0.63	0.50
6	Yes	0.65	0.65	0.65	0.65	0.34
7	Yes	0.66	0.66	0.66	0.66	0.22
8	Yes	0.68	0.68	0.68	0.68	0.16
9	Yes	0.69	0.69	0.69	0.69	0.16
10	Yes	0.70	0.70	0.70	0.70	0.11

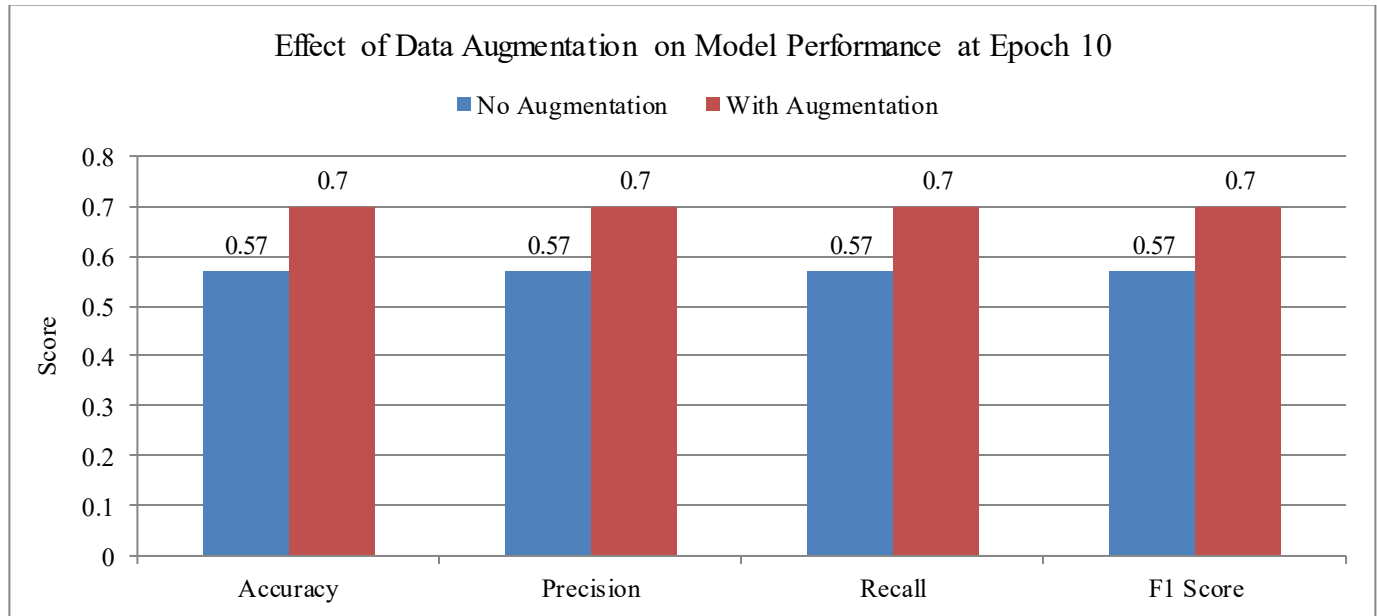


Fig. 12 Performance comparison at epoch 10 with (orange) and without (blue) data augmentation, highlighting superior results with augmentation

6. Error Analysis

6.1. Result Analysis for Fine-Grained Classification


In this section, we analyze the misclassifications made by the coarse-grained classification model, providing examples

of both correct and incorrect predictions. Table 9 shows the correct predictions, while Table 10 highlights the incorrect predictions, with justifications for why the model failed to classify them accurately.

Table 9. Example of correct prediction by coarse-grained classification model (1: Hostile, 0: non-hostile)

Post	Actual	Predicted
લુંટ સરકાર Translation: Loot government	1	1
સરકાર ખોટી છે આવા કાયદા નો હોય Translation: The government is wrong; laws like these should not exist.	1	1
આવા નાલાયક પુત્રોને કડક સજા મળવી જોઈએ Translation: Such incompetent sons should be given strict punishment.	1	1
પોલીસ એક સારું કામ કરે છે Translation: The police are doing a good job.	0	0
જેલ ભેગા કરો એટલે ખબર પડે Translation: If the jails are filled, then they will understand.	1	1
ઇસુદાન ભાઈ ને ખુબ ખુબ અભિનંદન આમ આદમી પાર્ટી જિંદાબાદ ખેડૂત ભાઈઓ જિંદાબાદ Translation: Congratulations to Isudan Bhai. Long live the Aam Aadmi Party, long live the farmer brothers!	0	0
સરકાર ગમે તે આવે મોંઘવારી નહીં ઘટે આ વાત સાચી છે Translation: No matter what the government does, inflation will not decrease – this is true.	1	1
હવે ખોટીના પેટના શુ થાવ શો Translation: Now, what will happen to the false narratives?	1	1
પાટીદાર સમાજ ફરી આંદોલન કરી જ દયો Translation: The Patidar community will once again start a movement.	1	1
ચુંટણી આવતા જ પાટીદાર સમાજ હીલોરે ચડવા ની તૈયારી કરી લાગે છે Translation: It seems that as the elections approach, the Patidar community is preparing to rise up.	1	1

Table 10. Example of incorrect prediction by coarse-grained classification model (1: Hostile, 0: non-hostile)

Post	Actual	Predicted	Justification
માણસ માણસાઈ ભૂલિયો સે એટલે આવું કરે બાકી ખાખી તો આજ સે ને કાલ નથી ભાઈ ભગવાન નો તો ડર રાખો વાલા Translation: When a person forgets humanity, this is what happens. Otherwise, the police are here today, not tomorrow, brother. Fear God.	1	0	This tweet's intention is directed at the police, but it is written in an indirect manner, causing it to be incorrectly classified.
બરાબર છે મેહુલભાઈ બુટલેગરો ને પકડે તો હપ્તા કોણ આપે Translation: Correct, Mehulbhai. If the bootleggers are caught, who will pay the extortion?	1	0	This tweet criticizes the Mehulbhai work, but the model cannot understand due to the implicit insult.
અંગ્રેજોની ભાગલા પાડો અને રાજ કરોની નિતી અપનાવી પાટીદાર આંદોલન તોડવા ગયેલી ભાજપને જડબાતોડ જવાબ મળ્યો છે Translation: The BJP, which went to break the Patidar movement, got a strong response after adopting the British strategy of dividing and ruling.	1	0	The model is unable to understand proverbs and quotes.
ચાટને વાલે લોગ દેશભક્તિ કી બાત ના કરે મુહ ખોલ કે  મુહ લે લે Translation: People who lick boots should not talk about patriotism. Open your mouth and take a stand.	1	0	This Hindi sentence is written in Gujarati, so the model cannot understand.
સાંસદ શ્રી @CRPaatil દ્વારા "हर घर तिरंगा, हर गाँव तिरंगा" જે અભિયાન અતંગતે ત્રિરંગો આજે મારા ઘરે પહોંચ્યો જે દેશભક્તિ સાથે જોડાયેલો છે અભિયાન ખૂબ સરસ બિરદાવા જેવું છે ખાસ કરીને	0	1	This sentence does not contain any unfair context, but the model is unable to understand due to code-mixed language (Hindi and Gujarati) being used.

સાંસદ શ્રી @CRPaatil ખૂબ ખૂબ અભિનંદન https://t.co/5txwfG14Vv Translation: Har Ghar Tiranga, Har Gaon Tiranga" campaign by MP @CRPaatil is great. The tricolour has reached my home today, symbolizing patriotism. The campaign is commendable, especially MP @CRPaatil. Congratulations!			
આ સરકાર ને કાઢવાનો ટાઇમ આવી ગયો છે જોગો Translation: It's time to remove this government, Jogo.	1	0	This tweet forces the model to incorrectly predict Non-hostile because it does not understand which government is being discussed and For what reason?
@Divya_Bhaskar @INC Gujarat @HardikPatel_ કોંગ્રેસ ની પોલ ખોલવાની જરૂર જ ક્યાં હતી? પોલ ખૂલ્લી જ હતી/છે. આમાં હાર્દિક નું તકસાધુપણું ખૂલી ગયું!!! Translation: @Divya_Bhaskar @INC Gujarat @HardikPatel_ Was there any need to expose the Congress? The truth was already out. In this, Hardik's opportunism has been revealed!	1	0	This tweet does not correctly classify due to an unseen word 'pol' being used. The model cannot understand an unseen word.

6.2. Result Analysis for Fine-Grained Classification

In fine-grained classification, each post is sorted into specific types, such as hate posts, fake posts, offensive posts, or defamation posts. Table 11 shows examples where the model made correct and incorrect predictions. This task is more difficult because some posts can fit into more than one category, or do not give enough context to understand clearly.

The model also faces trouble when posts refer to background events, include images, or use new slang or symbols.

Another challenge is the way abusive language keeps changing over a period of time, along with the use of mixed languages or informal writing styles, which makes classification more difficult.

Table 11. Examples of correct and incorrect predictions by a fine-grained classification model

Post	Actual	Predicted
જમ્મુ કાશ્મીર: અનંતનાગ એન્કાઉન્ટરમાં 2 આતંકવાદી ઠાર મરાયા, સર્ચ ઓપરેશન હજુ પણ શરૂ #Jammukashmir #CGNews Translation: Jammu Kashmir: In an encounter in Anantnag, 2 terrorists were killed. The search operation is still ongoing.	Offensive	Fake
રશિયા પર હજુ કેટલાક પ્રતિબંધ લગાવશે અમેરિકા Translation: USA will still impose some sanctions on Russia.	Fake	Offensive
આતંકવાદી #YasinMalik Translation: Terrorist #YasinMalik	Offensive	Offensive
ઉત્તર કોરિયાના પરમાણુ બોમ્બના પરીક્ષણની તૈયારીની ખબરથી અમેરિકા તણાવમાં #NorthHKorea #USNavy #NuclearTest https://t.co/1UKsHPtEqk Translation: The USA is under tension due to news of North Korea preparing for nuclear bomb tests.	Fake	Fake
મોરબીમાં જૂની અદાવતનો ખાર રાખી 3 શખ્સોએ યુવકને લમઘારી જાનથી મારી નાખવાની ધમકી આપી https://t.co/nPv5DbIY14 Translation: In Morbi, due to an old grudge, 3 individuals threatened a youth with a long knife.	Hate	Hate
યુનિફોર્મ પહેર્યો: હેડ કોન્સ્ટેબલે દુકાનદારને થપ્પડ મારી, ધમકી આપી https://t.co/canmK5uZem Translation: A Head Constable in uniform slapped a shopkeeper and threatened him.	Hate	Hate
પાટીદાર આંદોલન હિંસક બન્યું; મહેસાણામાં કફરું, મોબાઈલ ઈન્ટરનેટ સેવા સ્થગિત, આવતીકાલે 'ગુજરાત બંધ'નું એલાન The Patidar Movement turned violent; curfew was imposed in Mehsana, mobile internet services were suspended, and 'Gujarat Bandh' was announced for tomorrow.	Hate	Hate

The fine-grained classification task involves categorizing hostile posts into specific categories like hate, fake, offensive, and defamation posts. While our model performed fairly well in identifying some of these categories-such as posts related to violence, threats, or spreading false information-it also faced several challenges in identifying the correct category. At times, it struggled to separate closely related types of content, especially when the context was unclear, the meaning implied, or the post used mixed languages.

Misclassifications often happened when the model lacked background understanding, or when posts included slang, changing language patterns, or references to images or videos. Still, the results show that transformer-based models hold promise for this task in Gujarati. We need better-quality data, stronger context handling, and more support for the language's unique features and variations to improve further.

7. Discussion

In binary classification, it is evident from the experimental results that the fine-tuned Multilingual BERT (mBERT) model outperforms DistilBERT, GujaratiBERT, XLNet and DeBERTa, which are transformer-based models. The mBERT model for hostile post detection in the Gujarati language achieves 0.89 accuracy with a very good F1 score. The performance is optimized by Hyperparameter tuning with a learning rate of $1e-5$ and 20 to 25 epochs. It is interesting to note that different batch sizes have a minimal impact on results, indicating that the model is robust across different training configurations. In both coarse-grained and fine-grained tasks of hostile post detection, the data augmentation techniques helped to improve the model's performance, especially for a low-resource language like Gujarati. For coarse-grained classification, the accuracy increases from 0.88 to 0.93; for fine-grained classification, the accuracy increases from 0.57 to 0.70 after adding more varied training data. It is observed from the error analysis that the misclassifications are due to the use of proverbs and idiomatic expressions, the use of code-mixed language and emerging slang. The model finds difficulty in interpreting code-mixed language, particularly between Gujarati and Hindi and social media posts that contain criticism without hostile markers. These observations highlight the scope of improvement for current approaches. The challenges observed in fine-grained classification are intense due to the complexity of categorizing

posts into specific subtypes like hate posts, fake posts, offensive posts or defamation posts. The classification accuracy is also hampered due to insufficient contextual information, the inclusion of images and rapidly evolving abusive language. These issues suggest future research directions such as incorporating text with images, continuous updating of slang lexicons and code-mixed language in the language models. Overall, the results signify that the pretrained transformer models with hyperparameter tuning and with data augmentation offer a strong foundation for hostile post detection in the Gujarati language. However, addressing linguistic and cultural aspects remains essential for further improvement.

8. Conclusion

Hostile posts on social media platforms are growing very rapidly, which is a great matter of concern. It leads to mental stress, spreads hate, and disrupts healthy online interactions. There are automated hostile post detection solutions for widely spoken languages such as English. However, there is limited work in Indian regional languages such as Gujarati. This creates challenges in identifying harmful content and protecting users on social media. One of the major challenges is to have sufficient data for the low-resource languages. To address this issue, we have created the Gujarati Hostile Posts Detection (GuHPD) dataset, which consists of 14,800 social media comments annotated for various hostility categories. Data augmentation techniques were applied to train the model with sufficient data and diversity. Several transformer-based models were tested, and hyperparameters were tuned to enhance performance. The fine-tuned Multilingual BERT model showed promising results. After applying augmentation, the F1 scores improved to 0.93 from 0.89 and 0.70 from 0.57 for binary and fine-grained classifications, respectively. However, the model still faced difficulties in detecting code-mixed language, text with images and culturally implicit expressions. Our work provides a strong baseline for further research in this area, and future work can focus on expanding the dataset, refining class definitions and incorporating context-aware and domain-specific language models.

Data Availability

The Gujarati Hostile Posts Detection (GuHPD) dataset will be publicly available after the Ph.D. thesis defence.

References

- [1] Shazia Sajid et al., "Investigating how Cultural Contexts Shape Social Media Experiences and their Emotional Consequence," *Review of Education, Administration and Law*, vol. 7, no. 4, pp. 185-200, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Antoine Bordes, Léon Bottou, and Patrick Gallinari, "SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent," *Journal of Machine Learning Research*, vol. 10, no. 59, pp. 1737-1754, 2009. [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Hasan Beyari, "The Relationship Between Social Media and the Increase in Mental Health Problems," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, pp. 1-11, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [4] Tamara-Jade Kaz, "Myanmar: Facebook's Systems Promoted Violence Against Rohingya-Meta Owes Reparations," Amnesty International, 2022. [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Varad Bhatnagar, Prince Kumar, and Pushpak Bhattacharyya, "Investigating Hostile Post Detection in Hindi," *Neurocomputing*, vol. 474, pp. 60-81, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Raza Ali et al., "Hate Speech Detection on Twitter Using Transfer Learning," *Computer Speech & Language*, vol. 74, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeflal, "BERT-Based Ensemble Learning for Multi-Aspect Hate Speech Detection," *Cluster Computing*, vol. 27, no. 1, pp. 325-339, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Amit Praseed, Jelwin Rodrigues, and P. Santhi Thilagam, "Hindi Fake News Detection Using Transformer Ensembles," *Engineering Applications of Artificial Intelligence*, vol. 119, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Ramchandra Joshi et al., "Evaluation of Deep Learning Models for Hostility Detection in Hindi Text," *2021 6th International Conference for Convergence in Technology (I2CT)*, Maharashtra, India, pp. 1-5, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Anushka Sharma, and Rishabh Kaushal, "Detecting Hate Speech in Hindi in Online Social Media," *2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, Jaipur, India, pp. 1-5, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Mohit Bhardwaj et al., "HostileNet: Multilabel Hostile Post Detection in Hindi," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1842-1852, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Chander Shekhar et al., "Walk in Wild: An Ensemble Approach for Hostility Detection in Hindi Posts," *arXiv Preprint*, pp. 1-10, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Saurabh R. Sangwan, and M.P.S. Bhatia, "Denigrate Comment Detection in Low-Resource Hindi Language Using Attention-Based Residual Networks," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1-14, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Farhan Ahmad Jafri et al., "Uncovering Political Hate Speech During Indian Election Campaign: A New Low-Resource Dataset and Baselines," *arXiv Preprint*, pp. 1-5, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Angana Chakraborty, Subhankar Joardar, and Arif Ahmed Sekh, "Ensemble Classifier for Hindi Hostile Content Detection," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, pp. 1-17, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Taehyeon Kim et al., "Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation," *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 2628-2636, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Deepawali Sharma, Vivek Kumar Singh, and Vedika Gupta, "TABHATE: A Target-Based Hate Speech Detection Dataset in Hindi," *Social Network Analysis and Mining*, vol. 14, no. 1, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Arpan Nandi et al., "A Survey of Hate Speech Detection in Indian Languages," *Social Network Analysis and Mining*, vol. 14, no. 1, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Fatima-zahra El-Alami, Said Ouattik El Alaoui, and Nouredine En Nahnahi, "A Multilingual Offensive Language Detection Method Based on Transfer Learning from Transformer Fine-Tuning Model," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6048-6056, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah, "Leveraging Transformers for Hate Speech Detection in Conversational Code-Mixed Tweets," *arXiv Preprint*, pp. 1-12, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Benjamin Muller et al., "First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT," *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2214-2231, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Satyajit Kamble, and Aditya Joshi, "Hate Speech Detection from Code-Mixed Hindi-English Tweets Using Deep Learning Models," *arXiv Preprint*, pp. 1-6, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Arushi Sharma, Anubha Kabra, and Minni Jain, "Ceasing Hate with MOH: Hate Speech Detection in Hindi-English Code-Switched Language," *Information Processing & Management*, vol. 59, no. 1, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Tharindu Ranasinghe, and Marcos Zampieri, "Multilingual Offensive Language Identification for Low-Resource Languages," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1-13, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Tanmay Chavan et al., "A Twitter BERT Approach for Offensive Language Detection in Marathi," *arXiv Preprint*, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi, "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi," *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Dubai, United Arab Emirates, pp. 121-128, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [27] Onkar Litake et al., "Mono Versus Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition," *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pp. 607-618, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Prachi Shedge, Siddhi Kamalkar, and Deepa Gupta, "Hate Speech Detection in Marathi Tweets Using Stacked Deep Learning Models," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, pp. 639-650, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Arpan Nandi et al., "Combining Multiple Pre-Trained Models for Hate Speech Detection in Bengali, Marathi, and Hindi," *Multimedia Tools and Applications*, vol. 83, pp. 77733-77757, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Abhishek Velankar et al., "Hate and Offensive Speech Detection in Hindi and Marathi," *arXiv Preprint*, pp. 1-9, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Abhishek Velankar et al., "L3Cube-Mahahate: A Tweet-Based Marathi Hate Speech Detection Dataset and BERT Models," *Proceedings of the 3rd Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, Gyeongju, Republic of Korea, pp. 1-9, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Bhargav Chhaya et al., "SamPar: A Marathi Hate Speech Dataset for Homophobia, Transphobia," *International Conference on Speech and Language Technologies for Low-Resource Languages*, pp. 34-51, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] Muhammad Deedahwar Mazhar Qureshi et al., "Hate Speech Classification for Sinhalese and Gujarati," *Forum for Information Retrieval Evaluation (FIRE-Working Notes 2023)*, Goa, India, pp. 501-515, 2023. [[Google Scholar](#)] [[Publisher Link](#)]
- [34] Prasanna Kumar Kumaresan et al., "Dataset for Identification of Homophobia and Transphobia for Telugu, Kannada, and Gujarati," *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, pp. 4404-4411, 2024. [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Nikhil Narayan et al., "Hate Speech and Offensive Content Detection in Indo-Aryan Languages: A Battle of LSTM and Transformers," *arXiv Preprint*, pp. 1-15, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] Mounika Marreddy et al., "Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for Four Different NLP Tasks in Telugu Language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 1, pp. 1-34, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [37] Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N. Govaichelvan, "Tamil Offensive Language Detection: Supervised versus Unsupervised Learning Approaches," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1-14, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [38] Koyel Ghosh et al., "Transformer-Based Hate Speech Detection in Assamese," *2023 IEEE Guwahati Subsection Conference (GCON)*, Guwahati, India, pp. 1-5, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [39] G. Gnana Sai et al., "Enhancing Hate Speech Detection in Sinhala and Gujarati: Leveraging BERT Models and Linguistic Constraints," *Forum for Information Retrieval Evaluation (FIRE-Working Notes 2023)*, Goa, India, pp. 435-444, 2023. [[Google Scholar](#)] [[Publisher Link](#)]
- [40] Rezaul Haque et al., "Multi-Class Sentiment Classification on Bengali Social Media Comments Using Machine Learning," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 21-35, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [41] Raghad Alshaalan, and Hend Al-Khalifa, "Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach," *Proceedings of the 5th Arabic Natural Language Processing Workshop*, Barcelona, Spain, pp. 12-23, 2020. [[Google Scholar](#)] [[Publisher Link](#)]
- [42] Monil Gokani, and Radhika Mamidi, "GSAC: A Gujarati Sentiment Analysis Corpus from Twitter," *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada, pp. 129-137, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [43] Wikipedia, Gujarati Language, Wikipedia: The Free Encyclopedia, 2002. [Online]. Available: https://en.wikipedia.org/wiki/Gujarati_language
- [44] Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, vol. 1, pp. 4171-4186, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [45] K. Sreelakshmi, B. Premjith, and K.P. Soman, "Detection of Hate Speech Text in Hindi-English Code-Mixed Data," *Procedia Computer Science*, vol. 172, pp. 737-744, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [46] Rukhma Qasim et al., "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," *Journal of Healthcare Engineering*, pp. 1-17, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [47] Peluru Janardhana Rao et al., "An Efficient Methodology for Identifying the Similarity Between Languages with Levenshtein Distance," *International Conference on Communications and Cyber Physical Engineering 2018*, Hyderabad, India, pp. 161-174, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [48] Translation AI, Accelerate Global Growth with Quality Translation at Scale, Powered by Gemini, Google Cloud, 2025. [Online]. Available: <https://cloud.google.com/translate?hl=en>

- [49] Jonas Moss, "Measures of Agreement with Multiple Raters: Fréchet Variances and Inference," *Psychometrika*, vol. 89, no. 2, pp. 517-541, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [50] DistilBERT, Hugging Face, 2019. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/distilbert
- [51] Google-Bert/BERT-Base-Multilingual-Cased, Hugging Face, 2024. [Online]. Available: <https://huggingface.co/google-bert/bert-base-multilingual-cased>
- [52] Zhilin Yang et al., XLNet, Hugging Face, 2019. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/xlnet
- [53] Pengcheng He et al., DeBERTa, Hugging Face, 2018. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/deberta
- [54] RoBERTa, Hugging Face, 2019. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/roberta
- [55] L3cube-Pune/Gujarati-Bert, Hugging Face, 2022. [Online]. Available: <https://huggingface.co/l3cube-pune/gujarati-bert>
- [56] L3cube-Pune/Gujarati-Bert-Scratch, Hugging Face, 2022. [Online]. Available: <https://huggingface.co/l3cube-pune/gujarati-bert-scratch>
- [57] L3cube-pune/Gujarati-Sentence-Bert-Nli, Hugging Face, 2023. [Online]. Available: <https://huggingface.co/l3cube-pune/gujarati-sentence-bert-nli>
- [58] Budi Nugroho, and Anny Yuniarti, "Performance of Root-Mean-Square Propagation and Adaptive Gradient Optimization Algorithms on Covid-19 Pneumonia Classification," *2022 IEEE 8th Information Technology International Seminar (ITIS)*, Surabaya, Indonesia, pp. 333-338, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [59] Google/Muril-Base-Cased, Hugging Face, 2018. [Online]. Available: <https://huggingface.co/google/muril-base-cased>
- [60] L3cube-Pune/Marathi-Bert, Hugging Face, 2022. [Online]. Available: <https://huggingface.co/l3cube-pune/marathi-bert>
- [61] Google Bert /Bert-Large-Uncased, Hugging Face, 2024. [Online]. Available: <https://huggingface.co/google-bert/bert-large-uncased>