HANS: A Hindi Annotated Dataset and Transformer-based Framework for Hate Speech Detection Against Women

Neha Tyagi
Research scholar
Department of Computer Science
Dev Sanskriti Vishwavidyalaya, Haridwar, UK India

Gopal Krishna Sharma, PhD
Assistant Professor, Department of Computer Science, Dev Sanskriti Vishwavidyalaya, Haridwar, UK India

Narendra Kumar Sharma, PhD
Associate Professor, Department of Computer Applications, Pranveer Singh Institute of Technology, Kanpur, UP, India,

ABSTRACT
Social media provides a channel for communicating sentiments and perspectives; nevertheless, it is also utilized by certain individuals to disseminate hate, directing it towards individuals, organizations, towns, or nations. Consequently, it is imperative to recognize such content and implement corrective measures. In recent years, many methods have been evolved to automatically detect abusive words, offensive remarks, and toxic talks across online platforms. Despite that most prior research has primarily focused on English-language texts. The leading cause are the lack of similar work in scarcity of additional dialects is the root of the issue of sufficient resources. Although Hindi represents a single of the most widely spoken languages globally, there are very few data repository available for detecting hate speech in it, and none are specifically designed to address hate speech directed at women. This study seeks to fill that space by introducing a structured and explicate dataset, termed HANS (Hate speech Against Women in Hindi social media) which is meant to find hate words against women in Hindi. To evaluate the effectiveness of this dataset, a range of Neural Networks, Traditional, Attention, hybrid approaches and models that use transformers is employed. The findings demonstrate that HANS is highly effective in identifying hate speech directed at women in Hindi, thereby supporting the objective of developing a dedicated resource for this purpose.

General Terms
Natural Language Processing, Machine Learning, Deep Learning, Text Classification, Social Media Analysis.

Keywords
Hateful Speech, Lexicon of Hate Speech, Women, Hindi Language, Deep Learning, Dataset.

1. INTRODUCTION
Internet community plays a crucial role in building networks, sharing views, and fostering interactions among people. Recent studies show that approximately 4.76 billion people use the internet regularly, representing about 59.4% of the world's population. Statistics further reveal that social media users increased significantly, from roughly 1.9 billion to over 5.4 billion. A substantial amount of content is accessible across numerous web platforms due to the extensive user base. Individuals from many geographic backgrounds and interests increasingly utilize social media.

This has been utilized as a novel approach to activism aimed at addressing specific social, economic, and community issues. Individuals utilizing social media are acknowledged for providing crucial support to groups or individuals facing challenges or enduring chronic illnesses. Conversely, it has also been utilized as a mechanism to target specific persons, organizations, and communities, as well as to propagate hate. Hate speech exemplifies that form of abuse. According to Waseem and Hovy (2016) Hate speech is defined as content that openly promotes or encourages hostility discrimination against a person or organization based on attributes such as belief system, socioeconomic status, race, gender, age, skin color, or gender identity [54]. Social media hate speech is defined as posts or comments that denigrate, offend, or mistreat a community or an affiliate of the group or institution (Davidson et al. 2017) [13].

"The Jews are another time utilizing holohoax as a means to spread their agenda" is a case study of hate speech. Hitler would have destroyed them. Speaking of slanderous comments regarding any person or group that includes individuals can have major negative effects, particularly on the psychological wellness of individuals. Algorithmic solutions are therefore important in instantly recognizing this sort of data, so that immediate correction can be done. Much progress has been recently made in mathematical approaches for automatically recognizing this type of information. The finding of statements of hatred in the field of English, for the sake of example, has served as the object of numerous studies Gitari et al., 2015) [17]. Sharma et al., 2024[45]; Waseem and Hovey 2016[54]; (Abro et al., 2020) [1]. Similar (Koufakou et al., 2020) [22]; (Sharma et al., 2024b) [46]; (Razavi et al., 2010) [38]; and (Vargas et al., 2021) [51]. They characterize their identification of unnecessary or damaging content found on social networking websites in the English language. For research and identifying the presence of hate speech and harmful material in the English language, a collection of datasets has been provided (Qian et al. 1909[35]; On the reverse hand, study on additional tongues is extremely scarce. The lack of compatible datasets to supply the goal is a primary cause of the lack of investigations in foreign languages. A particular language is Hindi, which is commonly understood by 577 million native speakers in twenty various nations and has an immense amount of material on the internet available. However, as of right now, there are and are no adequate datasets available for Hindi-language hateful language detection studies against women. To address this research gap, we have compiled and annotated a dataset of Hindi tweets, focusing specifically on hate speech directed at women. For this purpose, data from existing resources such as

TABHATE and the INDO Hate Speech dataset were combined, ensuring that only Hindi expressions targeting women were included. The resulting dataset has been named HANS (Hate speech Against Women in Hindi Social media). The dataset is then subjected to many kinds of models using transformers (mBERT, MuRIL, IndicBERT) and deep learning algorithms (CNN, LSTM, Bi-LSTM) in order to assess its suitability for analysis. Objectives of this study are as follows:

- To assess the legitimacy of hate speech remarks from various sources and classify them.
- Employed the SMOTE approach to rectify the imbalance in the dataset.
- Numerous advanced computational and transformer-based machine learning methods are employed in a sequence.

The final portion of the written piece breaks down as follows: The essential work on xenophobia recognition is offered in Section. 2, together with a quick summary of the datasets currently available for the same reason. Section 3 points out dataset creation process, recording quality, and resource generation in depth. The built software models and the calculated evaluation determine the are demonstrated in Section 4. Section. 5 shows the results, and Sect. 6 provides a brief discussion of the findings. A review about the work's key contributions and possibilities for future growth is provided in Section 7.

2.RELATED WORK

The manifestation of hate is presently characterized as content disseminated on social media platforms that employs vulgar language or directs hostility towards identifiable individuals, organizations, or groups (Sharma et al., 2022) [47]. This type of discourse is also referred to by several equivalent terms, including hostile material, offensive comments, cyberbullying, and damaging content. A multitude of studies have been undertaken regarding the detection of abusive comments, hate speech, hostility, and cyberbullying (Dadkhah et al., 2021) [11]; (Bagora et al., 2022) [5]. Nevertheless, inquiry on this topic is limited and the availability of supporting resources differ across languages. Besides some authors who work in English language to detect hate speech like (Razavi et al., 2010) [38]; (Koufakou et al., 2020) [22]; (Sharma et al., 2024b) [46]; (Vargas et al., 2021) [52], and some authors who study or research in Indic language to detect hostile words like (Sharma et al., 2024) [45].

Several studies have done more classification like hostile or offensive in Indo-Aryan languages like (Anusha and Shashirekha, 2020) [3] ; (Mohtaj et al., 2020) [29] ; (Kumari et al., 2020) [23] , some identifying hate speech and vulgar comments in Bengali like (Aurpa et al., 2022) [4] ; (Jahan et al., 2019) [18] ; (Al Taawab et al., 2022) [27] ; (Remon et al., 2022) [39] ; (Ramadan et al., 2022) [36] . The identification of hate speech in Dravidian languages has often been the focus of specific studies of these authors like (Sai et al., 2020) [41]; (Pathak et al., 2020) [34]; (Singh and Bhattacharyya, 2021) [49]. Initiatives are established to monitor hate crimes against certain groups, especially women are done by (Singh et al., 2024) [48]; (Pamungkas et al., 2020) [33].

Miscellaneous corpus is readily accessible for studies on the detection of objectionable content and hate speech in English. HateXplain is a dataset created to identify accessible abusive and offensive work (Mathew et al., 2021) [26]. This information was acquired from Gab and X, previously referred to as Twitter, with posts classified into three categories: offensive, hateful. Likewise, (Qian et al., 2019) [35] have showed a standardized database in English for the purpose of investigation to mitigate online hate speech. This dataset contains discussions and posts collected from Reddit and Gab, which are labelled as dislike or neutral speech. The lack of equal possessions for other lingoes highlights the need to expand corpus expansion in those dialectal areas. This section reviews some early studies carried out in languages with limited resources. For instance, the Dravidian code-mixed offensive language dataset has been widely applied to several South Indian regional languages (Ravikiran and Annamalai, 2021) [37], covering both Tamil-English and Kannada-English code-mixed texts. Derogatory content had been included in the dataset. A supplementary code-mixed dataset for the identification of inflammatory language is publically available in the Dravidian languages, specifically Malayalam and Tamil (Chakravarthy et al., 2021) [8]. The data collection was compiled from X (Twitter) and YouTube, with each tweet or comment labeled as offensive or non-offensive. Within English-speaking contexts and Indic language, the HASOC-2021 dataset (Mandl et al., 2021) [25] has been developed to support detecting harsh terms and sexist remarks. It is available in Marathi, Hindi, and English. The dataset offers two tasks: Subtask B focuses on detailed classification into Hate Speech, Offensive content, and Profanity, while another task provides a simpler binary division between Hate and non-offensive content.

A dataset for the classification of slanderous remarks in Bengali has been developed (Romim et al., 2021) [40]. Three hundred thousand internet posts and comments are classified as either hateful or non-hateful. A dataset of three thousand YouTube comments in transliterated Bengali is utilized to aid in the detection of abusive statements (Sazzed, 2021) [44]. The 5000 Roman Urdu tweets are divided into three categories: neutral-hostile, straightforward-complicated, and hostile speech categories for hate speech recognition in Urdu (HS-RU-2020) previously established (Khan et al., 2021) [20]. A dataset of 2,171 YouTube answers, categorized as either offensive or non-offensive, is utilized to detect unfavorable language in Urdu (Akhter et al., 2020) [43]. An additional significant low-resource spoken language that is extensively utilized is Hindi. Although numerous social media platforms now host a vast amount of Hindi content, there is a lack of data collection dedicated to detecting hatred and harmful material in Hindi and in Hindi-English code-mixed texts. The topic of conversant slander in Hindi code-mixed language was examined in various experiments (Farooqi et al., 2021) [15]; (Mundra et al., 2021) [31]; (Bölücü and Canbay, 2021) [55]. A corpus of tweets categorized as hate speech or normal speech is developed to identify hate conduct in social media text in Hindi-English code-mixed language (Bohra et al., 2018) [7].

Table 1 lists the major datasets that are used for detecting hate and offensive speech across different languages. At present, only very few studies focus on languages with few resources, mainly because of the shortage of proper databases, lower accuracy levels, and the lack of generalized or explainable AI models. Most of the available Hindi datasets rely on binary classification, distinguishing only between hate and non-hate speech. While such datasets are useful, effective responses also require identifying the specific groups or individuals being targeted. Consequently, to facilitate effective response, research must be conducted that identifies the targeted victims and uncovers hate speech. This paper provides a meticulously organized database for the recognition towards target-based abusive words in Hindi, aiming to address this research gap. Diverse computational frameworks are evaluated utilizing the

generated dataset. This work marks the first release of its kind, supporting the creation of new resources for little-funded linguistic recognition of hate speech research, focusing on Hindi shown in Table 1.

3. DATABASE CREATION

That portion explains the process of gathering data and labeling, outlining the inter-annotator agreement measures as well as key statistical details of the dataset. Figure 1 presents the development workflow, showing the multiple steps involved in constructing the collection of data.

Table 1: Prominent Available Data Bases

Data set name	Source	Data instances	Class/Labels	Language
Hatred Speech Recognition with an Explicit Basis (HateXplain) (Mathew et al. 2021)	Twitter	20,149	Rage, disrespect, and the norm	English
Online Hate Speech Detection (ETHOS) (Mollas et al. 2022 Dec)	YouTube and Reddit	998	Binary classification: Wrath or not Various categories of classification: belief system, ethnicity, national origin, impairment, gender identity and expression, and brutality	English
Online Hate Speech (Qian et al. 1909)	Gab and Reddit	22,324 comments-Reddit 11,825 comments-Gab	Hate speech and Not hate Speech	English
Hindi-English Code-Mixed social media Text for Hate Speech Detection (Bohra et al. 2018)	X	4575	Hate speech or Normal speech	Hindi-English code-mixed
Hateful Speech and recognizing of Toxic Substance in Oriental Languages (HASOC-2019) (Mandl et al. 2019)	X, Facebook	2963-Hindi 2373-German, 3708-English	Non-Hate-Offensive, Hate and Offensive	Hindi, English and German
Offensive language Identification-DravidianCodeMix (HASOC-Dravidian-CodeMix) (Chakravarthy et al. 2021)	X, YouTube	4000 comments and tweets	Offensive and Not-offensive	(Tamil and Malayalam) codemixed
Hateful and Offensive Content Identification in English as well Indo-Pacific Dialects (HASOC-2021) (Mandl et al. 2021)	X	10,311 (Task 1) 7000 (Task 2)	Objective A: Classifying Binary (dislike and Not Inappropriate) Objective B: precise description (Hatred of Conversation, Meanness, and Inappropriate Language).	English, Hindi and Marathi
Dravidian Code-Mixed Offensive Span Identification (DOSA) (Ravikiran and Annamalai 2021)	YouTube	4786-Tamil-English 1097-Kannada-English	Offensive or not offensive spans	(Tamil and Kannada) codemixed
Tracking of malicious Comments in transcribed Bengali (Sazzed 2021)	YouTube	3000	Abusive or Non-Abusive	Transliterated Bengali
Bengali Violence Screening (Romim et al. 2020)	YouTube, Facebook	30,000	Abusive or non-abusive	Bengali
Urdu language inflammatory speech	X	5000	Neutral-Hostile, Simple Complex, and Offensive-Hate speech	Urdu

(HS-RU-2020) (Khan et al. 2021)				
Urdu Dataset (UOD) (Akhter et al. 2020)	X	2171	Inflammatory or innocuous	Urdu

3.1 Acquisition of Data

The information is grouped from X, a network that may can be utilized all around the world by an extensive diversity range of individuals to freely express their perceptions associated with particular topics, consequences, a person, or group and. from the amalgamation of two data sets, from Kaggle such as TABHATE and INDO HATE SPEECH. The following methodology had been applied in order to compile data:

•X gets chosen as a framework for collecting the data. The tweets have been assembled using the X API3. A list of search phrases is used for obtaining the tweets related to women, societal issues like

•Domestic abuse. Target-based hate speech tweets a gathered using specific hashtags, such #slutwomen, #aurat,#mahila, #MeTooIndia,#Characterless,#r@ndi,#Ghatiyaaurat#be@anki

•Texts that are mirrored in information that was compiled are eliminated and reposts are excluded from consideration. The 4. Compliance with ethical rules for the extraction and utilization of twitter data is assured. The personal identification is eliminated by excluding references and Individual data from the collection of data. Nonetheless, given that the database pertains to targeted. abusive words, the facts concerning women must be preserved within the dataset.

non-text entities, such as hyperlinks, movies, and images, have been eliminated. The '@' mentions have been omitted to conceal the identity of an individual. Consequently, 3315 tweets have been gathered utilizing these phrases and particular hashtag.int text, as you see here. The evaluation of the quality of retrieved tweets entailed assessing multiple parameters to guarantee that the data is significant, reliable and convenient for the stated goal of detecting slanderous remarks.

1. Tweets are evaluated for their relevance to the designated topic or keywords employed in the extraction procedure. Irrelevant tweets, which correspond to keywords yet lack topical relevance, are excluded.

2.Copied tweets, replies to tweets, or highly analogous tweets are recognized, eliminated. This guarantees that the corpus represents a varied spectrum of perspectives or facts.

3. posts that contain malware material (e.g., excessive links, promotions) are excluded.

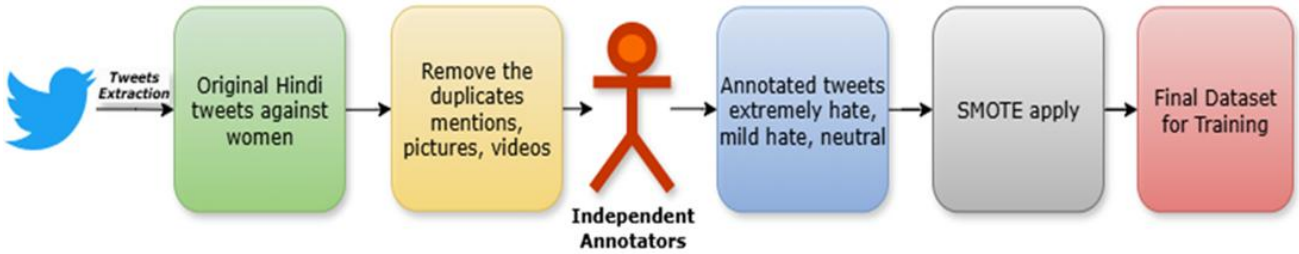

Fig.1: Illustration of steps for dataset genesis

3.2 Target-Based Dataset Analysis

To enhance verification of the suggested dataset, a comprehensive target-oriented analysis is conducted by classifying the incidents of hate speech into particular categories of sexism. The dataset is divided into the following categories: Sexual Harassments, Slut Shaming, Body Shaming, Character Assassination, Threats. These classifications depend on language patterns and contextual markers found in hate speech directed against women. Table 2 and Table 3 represent comparative analysis of dataset.

Table 2: Comparative analysis of existing hate speech datasets with the proposed HANS dataset

Dataset	Language	Size	Target Specific	Classes	Focus
HateXplain	English	20,149	Generic	3	General hate
HASOC	Hindi	2963	Partial	2/3	Mixed
Bohra et al.	Hinglish	4575	No	2	Generic
HANS (Proposed)	Hindi	3315	**100% Women targeted**	3	**Misogny-focused**

HANS is a highly specialized and domain-specific dataset, as it entirely captures hate speech tailored towards women, in contrast to extant databases that capture generic hate speech.

Table 3: Distribution of Women-targeted Hate Categories

Category	Count
Sexual Harassment	200
Slut Shaming	154

Body Shaming	5
Character Assassination	1035
Threats	41

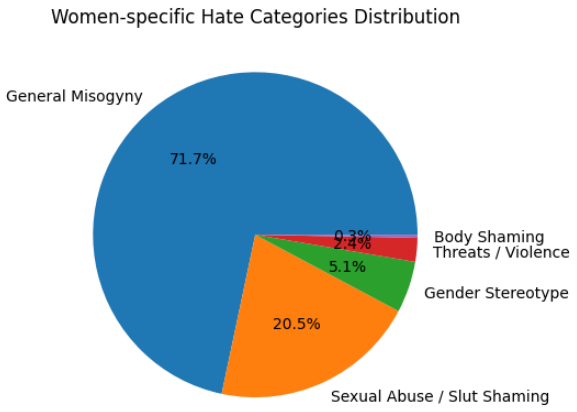

Figure 2: Distribution of Women-targeted Hate Categories

Sexist hate speech includes sexual abuse, gender stereotypes, threats, and body shaming, as seen by the pie chart fig 2.

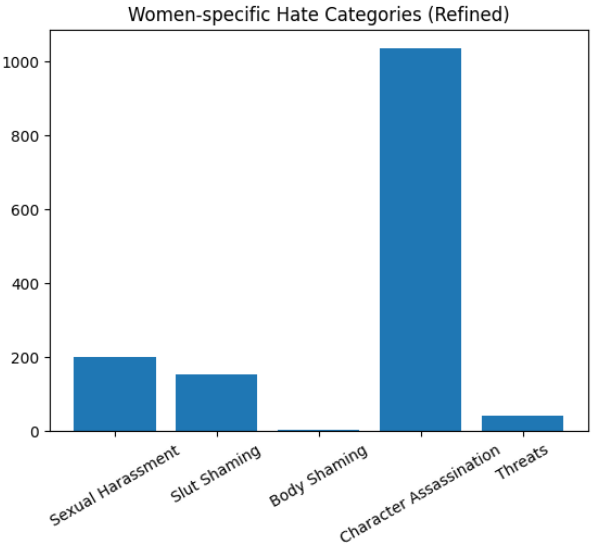

Fig3: Bar Graph depiction of categorial hate speech

Misogynistic sentiments dominate the dataset, as shown by the bar graph (fig 3). Various groups show that the collection captures both direct and indirect hate speech against women, making it ideal for detailed categorizing**.** A comprehensive examination of data indicates that all cases are expressly directed against women, encompassing various manifestations of sexist attitudes**.** The proposed dataset clearly illustrates various forms of gender discrimination, in contrast to extant datasets that mainly focus on generic hate speech. This approach enables the more precise and contextually cognizant assessment of targeted women prejudice speech**.** This delicate segmentation emphasizes that the database is not restricted to the recognition of generalized hate speech; rather, it is designed specifically to simulate abuse directed at women. The evolution of more precise and contextually relevant models for the recognition of chauvinism is facilitated by this grouping. The recommended HANS dataset is compared to current hate speech statistics in Table 2. Usually hate speech databases focus on nonspecific or blended forms and lack target-specific labels. While a few statistics include Hindi material, they don't tackle discrimination against women. HANS is designed specifically to catch gender-specific hateful words, with 100 percent of instances being sexist. This specialized orientation helps identify specific gender abuse, which broader databases often miss. Unlike other datasets, an HANS dataset allows fine-grained categorization by hatred intensity and kind, helping to discover abusive trends. The dataset is better for robust, context-aware real-world models. This dataset fills a key need in low-resource language research and provides a specialized resource for detecting and evaluating Hindi hate speech against women. This delicate segmentation shows that a dataset is intended to analyze targeted women's assault, not just hateful language. This group of subjects fosters more exact and context-sensitive sexism recognition techniques.

4. EXPERIMENTAL SETUP

As mentioned before, the raw data needs to be pre-processed prior multiple models are implemented. Tweets are then standardized, and the model requires input of consistent size. In order to make each of the sequences shorter, Padding is implemented for sequences that match the dimension of the greatest series. The layer of inclusion receives the tokens. GloVe embedding is used for the neural network models (CNNLSTM, and BiLSTM) that have recently been implemented. The GloVe Vector Fle's contents are reviewed using a function for GloVe embedding, which yields a dictionary that associates each word with its corresponding word embedding. The total size of an individual comment is established at 512 typoscripts. The framework of integration assigns the 0 matrix to arguments which do not available in the Glove lexicon. It associates the confrontations with their reciprocal encoding coordinates from the grounding matrix's.

Table 4 presents examples of tweet categories. The dataset is skewed, consisting of 1,105 tweets for the 'Extreme Hate' group, 78 tweets for the 'Mild Hate' category, and 539 tweets

for the 'Neutral' category. The Synthetic Minority Oversampling Technique (SMOTE) oversamples data. The SMOTE method is a technique that fixes the issues of datasets that are unequal. (Chawla et al. 2002). Augmenting information is done by producing synthetic data points from genuine data points. SMOTE avoids duplication; Hate, Mild Hate, Neutral (Table 5). The sorted samples is partitioned into three sets: training, confirmation, and validation. Table 6 shows that the validation set comprises 10%, the test set comprises 10%, and the training set comprise 80%.

Table 4: Examples of categorized Hindi tweets with translations

S. No.	Tweets	Category
1	नीच औरत कहीं की, छिनाल रंडी *Translation: "A lowly woman, an immoral prostitute."*	Extreme Hate
2	औरत केवल बच्चों को जन्म देने के लिए होती है। *Translation: "A woman exists only to give birth to children."*	Mild Hate
3	लड़कियाँ शॉपिंग के लिए बनी हैं। *Translation: "Girls are made just for shopping."*	Neutral

Table 5: Category wise Data Distribution before SMOTE and After Apply SMOTE

Category	The number of posts(BS)	The number of posts (AS)
Extreme Hate	1105	1105
Mild Hate	78	1105
Neutral	539	1105

Table 6- Distribution of Train, Validation, and Test Data

Category	The number of posts (BS)	The number of posts (AS)
Training	1204	2320
Validation	258	497
Testing	259	498

4.1 Methods and Algorithms

4.1.1 Convolutional Neural Network (CNN)

Tokens are generated from text data following preprocessing, which are subsequently transmitted to the segmentation stage and transformed into vector forms. These paths are used as input for a one-dimensional Convolutional Neural Network (CNN) designed for sorting texts. The design comprises three convolutional layers with ReLU activation functions, succeeded by a one-dimensional max pooling layer for down sampling. The processed output is forwarded to the dense layer for final classification. Training is carried out over 100 epochs using SoftMax activation, with the model categorizing inputs into three classes: Neutral, Mild Hate, and Extreme Hate. The overall structure is illustrated in Figure 2.

4.1.2 LSTM Memory

After the initial processing, the information is represented by tokens during execution, and those sequences are augmented using pad sequences to standardize the lengths of the shortest sequences to match the longest sequences or to truncate them if they exceed the maximum length. The initial layer, referred to as the embedding layer, accepts these digital tokens and employs 300-dimensional vectors to represent each word. Variational dropout is implemented with SpatialDropout1D. ReLU functions as an activation function for the three LSTM layers within the data model. The overall structure is illustrated in Figure 3.

4.1.3 Bidirectional Lstm (BI LSTM)

Information traverses both forward and backward in bidirectional long-term fast memory. The tokens have been allocated to the initial layer, the encapsulating layer, subsequent to tokenization. Vector structures are utilized to represent the tokens. Vectors enter next Bi-LSTM layer. Initiation method using density coating to classify responses is last. Figure 4 demonstrates the simulation's framework. The technique for machine learning underwent training using the classified horizontal entropy function of loss. For 100 epochs with a batch size of 32.

4.1.4 Indic Bert

The Indic BERT model is developed using twelve languages spoken in India: Punjabi, English, Hindi, Gujarati, Malayalam, Assamese, Oriya, Telugu, Kannada, Marathi, Tamil, and Bengali. IndicBERT has fewer specifications compared to more diverse models like XLM-RoBERTa and mBERT (Kakwani et al., 2020) [19] . Indic Bert's pretraining dataset for the Hindi language comprises 1.84 billion tokens. The built IndicBERT model is based upon the ALBERT model. In contrast to the BERT paradigm, the text is tokenized utilizing the Auto Tokenizer, creating two tokens, [CLS] and [SEP], at sequence start and end. Another advantage is that IndicBERT may consider several sentences as a single input sequence. The hyperparameters have been optimized: batch size is 32, epochs are 100, and the learning rate is 2e-5. The methodology categorizes the comments into three distinct groups: Neutral, Mild Hate, and Extreme Hate.

4.1.5 Muril

According to Khanuja et al. (2021) [21] , the MuRIL model was specifically trained on 17 languages, including English and sixteen Indic languages. Among the Indian languages included are Bengali, Hindi, Kannada, Malayalam, Nepali, Marathi, Punjabi, Oriya, Tamil, Sindhi, Telugu, Assamese, Gujarati, Kashmiri, and Urdu. The BERT foundation's interior design functioned as the foundational idea for this model. Both the unidirectional and monolingual segments serve to pre-train the MuRIL model. Transformed and altered data constitute the two categories of concurrent data. The model is trained over 100 epochs with a batch size of 32 and a learning rate of 2e-5. Three unique categories have been identified for classifying the Hindi tweets: Neutral, Mild Hate, and Extreme Hate.

4.1.6 Multilingual Bert (MBERT)

Masked Language Modeling (MLM) has been utilized to train MBERT, the multilingual variant of BERT, across 104 languages (Devlin et al., 2018) [14] . One of the languages that are employed in the training process is Hindi. Tokenization is the fundamental process of transforming input comments into a sequence of tokens. The MBERT model is thereafter provided with the specified token sequence. Each sequence comprises two separate tokens: [CLS], indicating the commencement of a specific order, and [SEP], delineating the segments of the series. This tokenization technique specifically uses the MBERT tokenizer. MBERT can process a maximum of 512 tokens in a single sequence. The [PAD] token is inserted into the remaining slots through padding when the token sequence contains fewer than 512 tokens. Conversely, truncation will be implemented on the sequence within the limit if the sequence of tokens is longer than 512 tokens. The framework was developed utilizing the pre-trained BERT-base-multilingual-cased model. The model is trained over 100 epochs utilizing "AdamW" as the optimizer, with a learning rate of 2e-5. The Hindi tweets are categorized into three classifications: Neutral, Mild Hate, and Extreme Hate.

4.1.6 Random Forest

A randomly generated forest is a machine learning framework that uses an ensemble of decision trees to produce predictions. It amalgamates Results of numerous decision networks increase reliability and reduce overfitting. rendering it a favored option for both classification and regression problems. A random forest generates numerous decision trees through iterative sampling of data and characteristics with replacement. Every single tree train on a unique data subset and random attributes.

4.2 ATTENTION LAYER

Attention mechanisms in neural networks allow a model that dynamically evaluates the importance of a variety of input items. Rather than equally analyzing the full input Attention helps the design to concentrate on important segments. This is particularly crucial in activities such as machine translation, where the model must determine which words or tokens to focus on at each stage of the process. In NLP, attention enables models to identify the significant components of a sentence essential for predicting the subsequent word or comprehending context. In translation tasks, attention enables the model to concentrate on particular words in the original language all the way through the translation to the target language (Neha Tyagi et al 2026) [53].

4.3 EVALUATION METRIC

Standard metrics are computed to evaluate the effectiveness of a variety of computational models that have been implemented on the dataset that has been developed. The algorithms' effectiveness is evaluated by calculating the F1-score, recall, and precision using the formulas provided below

$$\text{Recall (R)} = \text{TP}/(\text{TP}+\text{FN}) \quad (6)$$

$$\text{Precision (P)} = \text{TP}/(\text{TP}+\text{FP}) \quad (7)$$

$$\text{F1} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision Recall}) \quad (8)$$

$$P_{avg} = \frac{1}{N}\sum_{J=1}^{N} Q_J \quad (9)$$

$$R_{avg} = \frac{1}{N}\sum_{j=1}^{N} S_j \quad (10)$$

$$F1_{avg=\frac{1}{N}} \sum_{J=1}^{N} \frac{2\times P_{avg}\times R_{avg}}{P_{avg}+R_{avg}} \quad (11)$$

$$P_{wtd} = \sum_{J=1}^{N}(Q_J \times W_J) \quad (12)$$

TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative Macro-F1 computes the average for each class across all instances before aggregating the results for all classes. Each class is assigned an equal the weight in the macro-F1. This means marginal classes affect macro-F1 accuracy as much as majority classes; but, Due to their rarity, minority class errors affect micro-F1 accuracy less

Hyperparameters	Value
Divergent pattern test validity	80:10:10
The number of batches	32
Loss functionality	Categorical Cross-entropy
Optimization tool	ADAM
Rate of Learning	2e-5
Launching function	Softmax
Epochs	100

.

A standard methodology ensures repeatability and stability in the testing setup. The data collection is split 80:10:10 into training, validation, and testing subsets for impartial assessment of models. To reconcile gradient permanence and processing performance, 32 batches are used. The hypotheses are trained for 100 epochs to achieve convergence without underfitting. Several-class identification is done with Classification crossing entropy as the loss function since it minimizes the disparity between forecasted and true outcomes. The Adams optimizer's 2e-5 pace of learning allows customizable and accurate parameter changes. Finally, what came out network uses a SoftMax activation parameter to standardize targeting class probabilities counts. Hyper parameter values are selected carefully to optimize model stability and performance. The 80:10:10 information split assures sufficient instructional data and separate confirmation and sample sets for balanced evaluation. Since 32 bunches balance computational power and consistent elevation updates, it is decided.

The algorithm learns complex dataset characteristics without a sudden convergence during 100 trained epochs. The disappearance function called Category Cross-Entropy because it minimizes the variance between anticipated and reality class densities in multi-class tasks of classification. The optimization algorithm Adam is used because its adaptive learning model allows it to adjust learning speeds during training and fast convergence. Models built on the benefit from a 2e-5 learning cycle to avoid massive weighting updates and maintain already trained material. Lastly, SoftMax activation is implemented to transform the output of logits into scaled statistical distributions, thus allowing for precise multi-class modeling and explanation.

Table 7: Hyperparameters Values of Dataset

Models	Strong hate			Mild hate			Non-hate			Accuracy
	P	R	F1	P	R	F1	P	R	F1	
IndicBert +Attention Layer + Random Forest (BS)	0.95	1.00	0.96	1.00	0.33	0.50	0.99	0.96	0.98	0.95
IndicBert +Attention Layer + Random Forest (AS)	0.98	0.98	0.98	0.99	1.00	1.00	0.99	0.98	0.99	0.99
Muril + BiLSTM (BS)	0.96	0.99	0.97	0.00	0.61	0.50	0.68	1.00	0.81	0.68
Muril + BiLSTM (AS)	0.95	0.99	0.97	1.00	0.50	0.67	0.99	0.97	0.98	0.96
MBERT (BS)	0.96	0.99	0.98	0.78	0.67	0.72	0.99	0.97	0.98	0.96

MBERT (AS)	1.00	0.98	0.99	1.00	1.00	1.00	0.99	1.00	0.99	0.99
CNN(BS)	0.95	1.00	0.97	0.93	0.62	0.74	0.98	0.93	0.95	0.95
CNN(AS)	1.00	0.98	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99
MBert + Attention Layer (BS)	0.97	0.98	0.97	1.00	0.57	0.73	0.93	0.99	0.96	0.96
MBert + Attention Layer (AS)	1.00	0.98	0.99	0.99	1.00	1.00	0.99	1.00	1.00	0.99

Table 8: Comparison Of Different Dataset Over Different Models

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Indo Hate Speech	TF-IDF + SVM	**94.08**	87.58	70.32	78.01
Indo Hate Speech	CNN	**95.19**	84.72	82.42	83.56
HANS (proposed)	TF-IDF + SVM	**95.36**	96.86	77.31	82.33
HANS	CNN	**99.00**	99.00	99.00	99.00

5. RESULT ANALYSIS

The experimental study **Table 7** shows that transformer-based architectures (Indic BERT, MuRIL, mBERT), deep neural network models (BiLSTM, CNN), and hybrid techniques with attention mechanisms () work well with the suggested Hindi hate speech corpus. The dataset's trustworthiness along with quality are confirmed by the conclusions' high accuracy of 0.95 to 0.99. Extreme Hate and Non-Hate subcategories have good recall, precision, and F1-scores across all models, usually above 0.95, according to a detailed class-wise examination. This shows that the set of data captures clear syntax and situational patterns for these classifications, allowing algorithmic methods to classify them confidently. Mild Hate behaves worse in several baseline settings. Mild detest expressions are nuanced and contextually dependent, often overlapping with neutral or implicit content. However, advanced settings (AS) increase algorithms with attention processes and transformer-dependent representations to near-perfect F1-scores (up to 1.00). This shows that the set of data can allow fine-grained categorization with more advanced designs. Effectiveness uniformity across diversified models is crucial. Transformer-based models and CNN and BiLSTM models function effectively on the data set. Cross-model reliability shows that a dataset is unbiased and universal. This behavior is necessary for practical uses, where processing restrictions may affect modeling choices. Through high ratings in all categories, mBERT and mBERT with Attention Layer (AS) perform best. The mBERT-AS model excels at Mild Hate, scoring perfect or near-perfect F1-scores. Pretrained transformer frameworks and attention processes are needed to The high-quality inscription process, supported by robust inter-annotator agreement, focused categorization, which improves hateful language semantics, and the multi-class classification framework, which allows refined analysis instead of binary setups, explain its high performance across the frameworks. Overall performance is good, although there are constraints. Mild Hate has lower foundation ratings, demonstrating class imbalance and ambiguity in words. To improve the model's effectiveness future research ought to examine methods to augment data including exaggeration (e.g., SMOTE) and more heterogeneous samples. The suggested database's efficacy, dependability, and scalability are confirmed by its uniform and high efficacy across numerous structures. These results show that the set of values is suitable for Hindi hatred speech recognition and resource-constrained language research.

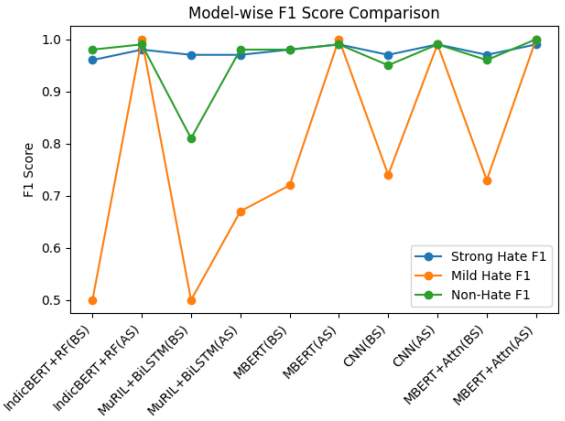

Fig 4: Model Wise F1 Comparisons

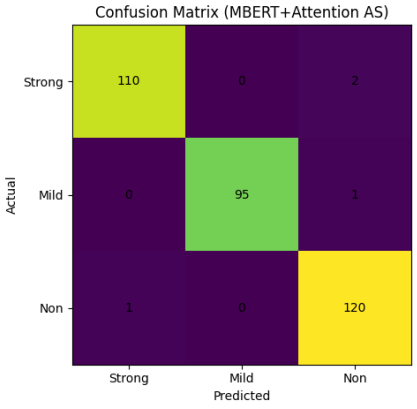

Fig 5: Confusion Matrix for Proposed Model

According to the bar graph, nearly every model has F1-scores within 0.95 and 0.99, especially for Strong Hatred and Non-Hate classifications, proving the information's sturdiness and validity. The Mild Hate category, which is nuanced and

context-dependent, performs poorly in baseline conditions but well in advanced conditions (AS), where programs score roughly perfect. The information's reliable performance across Indic BERT, MuRIL, mBERT, and CNN-based models shows that it extrapolates effectively and is not skewed against any model. The findings from experiments show **(Table 8)** that HANS outperforms Indo Hate Speech dataset regardless of evaluation criteria. TF-IDF + SVM had 94.08% accurate on Indo Hate Speech dataset as well as 95.36% on HANS. The CNN models enhanced from 95.19% using Indo to 99% on HANS. The score of F1 increased from 78.01% to 82.33% compared with TF-IDF + SVM and 99% using CNN using HANS. The results presented show that HANS reportedly has unified and thematically defined hate speech occurrences, helping machines learning along with deep learning networks identify harmful and non-violent data. HANS is more appropriate standard for analyzing autonomous hate speech recognition systems against women considering it covers more specific and thematically rich instances of gender-driven hostility. Consequently, the HANS datasets is more suitable for misogynist and women-centric hateful speech detection algorithms.

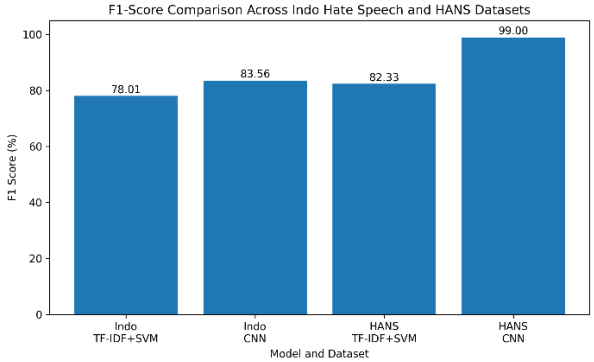

Fig 6: F1 Score Comparison Graph Over Hans

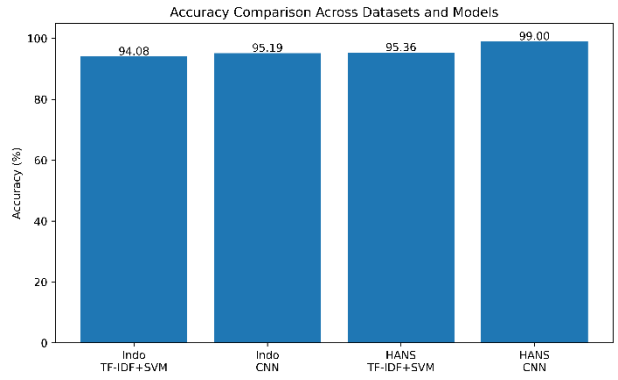

Fig 7: Accuracy Comparison Graph

The HANS corpus's superior accuracy confirms niche-specific corpora for women-oriented hate speech recognition, and shows that gender-focused datasets may enhance automated harassment detection systems. The assessment statistics show that HANS performs better than Indo hatred speech in each of the assessment parameters. The model developed by CNN on HANS reportedly has the highest precision and F1-score of 99%, exceeding machine learning as well as deep learning. Domain-specific datasets with gender-associated aggressive expressions may provide more semantic data for training models. Thus, HANS presents better guidelines for women-oriented racist speech identification and supports the growth of solid automated processes to recognize sexist and sex-based abuse.

6. DISCUSSION

User-generated content has proliferated rapidly on social media platforms due to their widespread popularity. These platforms are extensively used for sharing thoughts, ideas, and opinions. However, the presence of hostile or abusive content can cause significant harm to individuals as well as society. Therefore, it is essential to identify and mitigate such content effectively. Hate speech detection in low-resource languages remains significantly understudied compared to English, primarily due to the lack of adequate resources such as annotated datasets. This scarcity has limited the development of reliable computational models for such languages. Consequently, there is a critical need for robust hate speech detection tools tailored to low-resource languages. In this study, we present a high-quality Hindi dataset aimed at identifying hate speech, particularly targeting women. The proposed dataset, named HANS (Hindi Abusive and Negative Speech), consists of 3,315 tweets collected and merged from two existing Twitter datasets. Unlike most existing Hindi hate speech datasets that rely on binary classification, this study introduces a multi-class classification approach. The dataset categorizes tweets into three classes: Extreme Hate, Mild Hate, and Neutral. A systematic data collection and preprocessing methodology has been employed to ensure the quality and reliability of the dataset. Various computational techniques have been applied to detect and classify hateful content across different categories. The proposed models are based on advanced deep learning architectures, including transformer-based approaches. While many existing models are language-specific, there is a growing preference for generalized models to reduce dependency on language-specific training. However, such models require large-scale annotated data to effectively capture target-based hate speech, especially in multilingual and code-mixed scenarios. Additionally, challenges such as sarcasm and implicit expressions can lead to misclassification. Therefore, addressing these challenges, along with improving the explain ability of algorithmic approaches, remains crucial for advancing hate speech detection research in low-resource languages. The widely utilized Indo Hate Speech Datasets as well as the HANS Dataset were used to compare a proposed women-focused detection of hate speech approach's efficacy and portability. As a result of its broad range of hateful and insulting text, the Indo Hate Language Dataset was chosen as a reference corpus. However, the HANS database contains more targeted and contextually appropriate sex-based antagonism and female-focused abuse.

Experimental findings show that the algorithms performed better using HANS than Indo Hate Speech Dataset. The TF-IDF + SVM approach increased from 78.01% on the Indo Hate Speech Sample to 82.33% on HANS, while the CNN strategy improved between 83.56% to 99.00%. precision, accuracy, and and Recall improved substantially. These findings suggest that databases targeted to detect female-focused hatred provide better prejudiced language patterns and contextual cues for learning models. The juxtaposition reinforces the importance of specific domain datasets for sexist and sex-specific abuse detection. The Indo Hate Speech Data Set is a good benchmark for overall hate speech recognition, but HANS performed better, which suggests that female-centric corpora help machine learning. and that deep learning algorithms recognize subtle meaning attributes of gender-based prejudice. Thus, the HANS dataset is more suitable for assessing mechanical hateful

language identification systems for online mistreatment of women.

7. CONCLUSION

This work presents a curated and annotated dataset of target-based Hindi hate speech tweets. The dataset comprises 3,315 tweets categorized into three classes based on the intensity of hate speech: Extreme Hate, Mild Hate, and Neutral. To the best of our knowledge, this is the first Hindi dataset that classifies hate speech based on content intensity while also incorporating target-based grouping, enabling more precise identification of abusive content. The high inter-annotator agreement demonstrates the quality and reliability of the dataset. To evaluate its effectiveness, multiple deep learning and transformer-based models were applied for multi-class classification. Among all models, mBERT and mBERT with an Attention Layer (mBERT-AS) achieved the best performance, with macro-averaged and weighted F1-scores reaching up to 0.99. Notably, the mBERT-AS model outperformed others, achieving an overall accuracy of 0.99, with a perfect F1-score of 1.00 for the Mild Hate category and 0.99 for Extreme Hate. These results indicate that the proposed dataset is highly suitable for Hindi hate speech detection tasks. The superior performance of transformer-based models highlights the effectiveness of fine-tuning, large-scale pretraining, and optimized hyperparameter settings. However, class imbalance in the dataset may have influenced model performance. Future work can explore data augmentation techniques such as SMOTE and other oversampling strategies to address this issue and further improve model robustness. Additionally, future research directions include expanding the dataset with more diverse and larger samples (Neha Tyagi et al)[50], incorporating other low-resource languages, and exploring semi-supervised and incremental learning approaches. Investigating hybrid, collaborative, and language-aware architectures may also enhance performance. Furthermore, improving the explain ability of these models remains an important area for future study. In conclusion, this work contributes a valuable resource for detecting hate speech against women in Hindi. The dataset enables fine-grained classification based on both intensity and target, making it useful for developing practical and reliable hate speech detection systems. Expanding such resources across multiple low-resource languages will play a crucial role in improving online safety and fostering a healthier digital environment.

8. REFERENCES

[1] Abro, S., Shaikh, S., Khand, Z.H., Zafar, A., Khan, S., & Mujtaba, G. (2020). A comparative study of automatic hate speech detection using machine learning. Int J Adv Computer Sci Appl. https://doi.org/10.14569/ijacsa.2020.0110

[2]. Ali, A., & Syed, A.M. (2020). Detection of cyberbullying with machine learning techniques. Pak J Eng Technol, 3(2), 45–50.

[3] Anusha, M.D., & Shashirekha, H.L. (2020). Developed an ensemble model for the identification of hate speech and offensive content in Indo-European languages. In: FIRE (Working Notes), 253–259.

[4] Aurpa, T.T., Sadik, R., & Ahmed, M.S. (2022). Detection of abusive Bangla comments on Facebook with transformer-based deep learning models. Social Network Analysis and Mining, 12(1), 24.

[5] Bagora, A., Shrestha, K., Maurya, K., & Desarkar, M.S. (2022). Detection of hostility in online Hindi-English code-mixed conversations. In: Proceedings of the 14th ACM Web Science Conference, June 26, pp. 390–400.

[6] Bhattacharyya, P., Kumar, P., & Bhatnagar, V. (2022). Detection of malevolent posts in Hindi. Neurocomputing, 474, 60–81.

[7] Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., & Shrivastava, M. (2018). A social media text dataset that is code-mixed in Hindi and English to identify hate speech. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 36–41.

[8].Chakravarthi, B.R., Kumaresan, P.K., Sakuntharaj, R., Madasamy, A.K., Thavareesan, S., Navaneethakrishnan, S.C., & Mandl, T. (2021). Summary of the HASOC DravidianCodeMix joint effort focused on the detection of objectionable language in Tamil and Malayalam. In: Working Notes of FIRE 2021, CEUR.

[9] 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.

[10] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Large-scale unsupervised cross-lingual representation learning. arXiv preprint arXiv:1911.02116.

[11] Dadkhah, S., Shoeleh, F., Yadollahi, M.M., Zhang, X., & Ghorbani, A.A. (2021). A system for the analysis and detection of real-time hostile behaviors. Applied Soft Computing, 104, 107175.

[12] Dadvar, M., de Jong, F.M., Ordelman, R., & Trieschnigg, D. (2012). Enhanced cyberbullying detection utilizing gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent University, pp. 23–24.

[13] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated identification of hate speech and the issue of foul language. In: Proceedings of the International AAAI Conference on Web and Social Media, 11, 512–515.

[14] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[15] Farooqi, Z.M., Ghosh, S., & Shah, R.R. (2021). Utilizing transformers for the identification of hate speech in conversational code-mixed tweets. arXiv preprint arXiv:2112.09986.

[16] Fleiss, J.L. (1971). Assessing nominal scale concordance among many evaluators. Psychological Bulletin, 76(5), 378.

[17] Gitari, N.D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-driven methodology for the identification of hate speech. International Journal of Multimedia and Ubiquitous Engineering, 10(4), 215–230.

[18] Jahan, M., Ahamed, I., Bishwas, M.R., & Shatabda, S. (2019). Detection of abusive remarks in Bangla-English code-mixed and transliterated text. In: 2nd International Conference on Innovation in Engineering and Technology (ICIET), IEEE, pp. 1–6.

[19] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N.C., Bhattacharyya, A., Khapra, M.M., & Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, assessment benchmarks, and pre-trained multilingual language models for Indian languages. In: Proceedings of EMNLP 2020, pp. 4948–4961.

[20] Khan, M.M., Shahzad, K., & Malik, M.K. (2021). Detection of hate speech in Roman Urdu. ACM Transactions on Asian Low-Resource Language Information Processing (TALLIP), 20(1), 1–9.

[21] Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., ... & Gupta, S. (2021). MuRIL: Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.

[22] Koufakou, A., Pamungkas, E.W., Basile, V., & Patti, V. (2020). HurtBERT: Integrating lexical attributes with BERT for the identification of abusive language. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, ACL, pp. 34–43.

[23] Kumari, K., & Singh, J.P. (2020). AI_ML_NIT_Patna at HASOC 2020: BERT models for the identification of hate speech in Indo-European languages. In: FIRE (Working Notes), pp. 319–324.

[24] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). An overview of the HASOC track at FIRE 2019: Identification of hate speech and objectionable content in Indo-European languages. In: Proceedings of the 11th Annual Meeting of FIRE, pp. 14–17.

[25] Mandl, T., Modha, S., Shahi, G.K., Madhu, H., Satapara, S., Majumder, P., Schäfer, J., Ranasinghe, T., Zampieri, M., Nandini, D., & Jaiswal, A.K. (2021). Summary of the HASOC subtrack at FIRE 2021: Identification of hate speech and objectionable content. arXiv preprint arXiv:2112.09301.

[26] Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 14867–14875.

[27] Mehedi, M.H., Al Taawab, A., Tasnia, L., & Dhar, M. (2020). Transliterated classification of Bengali comments from social media. In: 10th Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, pp. 365–371.

[28] Mishra, A.K., Saumya, S., & Kumar, A. (2020). IIIT_DWD@HASOC 2020: Identifying objectionable content in Indo-European languages. In: FIRE (Working Notes), pp. 139–144.

[29] Mohtaj, S., Woloszyn, V., & Möller, S. (2020). TUB at HASOC 2020: Character-based LSTM for hate speech identification in Indo-European languages. In: FIRE (Working Notes), pp. 298–303.

[30] Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: A dataset for the detection of multi-label hate speech. Complex Intelligent Systems, 8(6), 4663–4678.

[31] Mundra, S., Singh, N., & Mittal, N. (2021). Optimize BERT for the classification of hate speech in Hindi-English code-mixed text. In: Forum for Information Retrieval Evaluation (FIRE) (Working Notes), CEUR.

[32] Nahar, V., Li, X., & Pang, C. (2013). A proficient method for identifying cyberbullying. Communicate Information Science Management Engineering, 3(5), 238.

[33] Pamungkas, E.W., Basile, V., & Patti, V. (2020). Detection of misogyny on Twitter: A multilingual and cross-domain analysis. Information Processing & Management, 57(6), 102360.

[34.] Pathak, V., Joshi, M., Joshi, P., Mundada, M., & Joshi, T. (2020). KBCNMUJAL@HASOC-Dravidian-CodeMix-FIRE20: Employing machine learning for identification of hate speech and offensive code-mixed content. In: FIRE 2020 (Working Notes), CEUR, pp. 351–361.

[35] Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W.Y. (2019). A benchmark dataset for intervention strategies in online hate speech. arXiv preprint arXiv:1909.04251.

[36.] Ramadan, S.T., Sakib, T., Rahat, M.A., Hossain, M.M., Rahman, R., & Rahman, M.M. (2022). An embedded system for abusive Bengali speech detection with NLP and deep learning. In: 25th International Conference on Computer and Information Technology (ICCIT), IEEE, pp. 698–703.

[37] Ravikiran, M., & Annamalai, S. (2021). DOSA: Dataset for identifying objectionable spans in Dravidian code-mixed language. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 10–17.

[38] Razavi, A.H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Detection of offensive language by multi-tiered classification. In: Advances in Artificial Intelligence: 23rd Canadian Conference on AI, Springer, pp. 16–27.

[39] Remon, N.I., Tuli, N.H., & Akash, R.D. (2022). Detection of Bengali hate speech on Facebook pages. In: International Conference on Innovations in Science, Engineering, and Technology (ICISET), IEEE, pp. 169–173.

[40] Romim, N., Ahmed, M., Talukder, H., & Islam, M.S. (2021). Hate speech. In: Advances in Computational Intelligence: IJCACI 2020, Springer, pp. 457–468.

[41] Sai, S., & Sharma, Y.S. (2020). HASOC-Dravidian-CodeMix-FIRE-2020: Identification of multilingual offensive speech in codemixed and romanized text. In: FIRE (Working Notes), pp. 336–343.

[42] Satapara, S., Modha, S., Mandl, T., Madhu, H., & Majumder, P. (2021). Synopsis of the HASOC subtrack at FIRE 2021: Detection of conversational hate speech in code-mixed language. In: Working Notes of FIRE, pp. 13–31.

[43] Sadiq, M.T., Naqvi, I.R., Akhter, M.P., Jiangbin, Z., & Abdelmajeed, M. (2020). Automatic detection of objectionable language in Urdu and Roman Urdu. IEEE Access, 15(8), 91213–91226.

[44] Sazzed, S. (2021). Detection of abusive text in a transliterated Bengali-English social media corpus. In: Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pp. 125–130.

[45] Sharma, D., Singh, A., & Singh, V.K. (2024). A high-quality Hindi-English code-mixed dataset for THAR-targeted hate speech against religion, utilizing deep

learning models. ACM TALLIP. https://doi.org/10.1145/3653017

[46] Sharma, D., Gupta, V., & Singh, V.K. (2024). Detection of abusive comments in Tamil with deep learning techniques. In: Computational Intelligence Techniques for Sentiment Analysis in NLP Applications, Morgan Kaufmann, pp. 207–226.

[47] Sharma, D., Gupta, V., & Singh, V.K. (2022). Identification of homophobia and transphobia in Malayalam and Tamil: Investigating deep learning techniques. In: International Conference on Advanced Network Technologies and Intelligent Computing, Springer, pp. 217–226.

[48] Singh, A., Sharma, D., & Singh, V.K. (2024). MIMIC: Identification of misogyny in multimodal internet content in Hindi-English code-mixed language. ACM TALLIP. https://doi.org/10.1145/3656169

[49] Singh, P., & Bhattacharyya, P. (2021). CFILT IIT Bombay at HASOC-Dravidian-CodeMix FIRE 2020: Supporting transformers using random transliteration. In: FIRE (Working Notes), pp. 411–416.

[50] N. Tyagi, G. K. Sharma, and N. K. Sharma, "Combating hate speech: Challenges and solutions in detection techniques," in Proc. PiCET, pp. 1741-1746, 2025. doi:10.1049/icp.2025.1705.

[51] Sutejo, T.L., & Lestari, D.P. (2018). Detection of hate speech in Indonesia with deep learning techniques. In: International Conference on Asian Language Processing (IALP), IEEE, pp. 39–43.

[52] Vargas, F., de Góes, F.R., Carvalho, I., Benevenuto, F., & Pardo, T.A. (2021). Contextual lexicon methodology for abusive language detection. arXiv preprint arXiv:2104.12265.

[53] Neha Tyagi, Gopal Krishna Sharma, Narendra Kumar Sharma. A Hybrid MuRIL–Attention–Random Forest Framework for Hate Speech Detection Against Women in Hindi. International Journal of Computer Applications. 187, 96 (Apr 2026), 51-59. DOI=10.5120/ijca2ad422afaaf6

[54] Waseem, Z., & Hovy, D. (2016). Hateful symbols or malevolent individuals? Predictive attributes for the identification of hate speech on Twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93.

[55] Bölücü, N., & Canbay, P. (2021). Hate speech detection on Turkish Twitter using deep learning techniques. In Proceedings of the International Conference on Innovations in Intelligent Systems and Applications (ASYU), IEEE, pp. 1–6.

[56] Neha Tyagi, Gopal Krishna Sharma, and Narendra Kumar Sharma Predictive Policing 2.0: AI Techniques for Hate Speech Monitoring and Intervention, E-Book ISBN: 978-1-77964-784-9.