IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review of Hate Speech Detection using Machine Learning Algorithm

Mrs. Preeti V. Sarode
Department of Computer Science and Information Technology
K. M. Agrawal College, Kalyan,
Maharashtra, India.

Harshali B. Patil
Department of Computer Science
Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya, Jalgaon,
Maharashtra, India.

ABSTRACT

In this digital era, social media is a popular & powerful tool to communicate digitally with each other. This daily communication generates the massive amount of electronic data on web. Processing this huge data is a challenging task. Hence social media data processing is gaining more focus. Hate speech detection is one of the important parts of social media data processing. This paper presents the review of hate speech detection systems developed using machine learning techniques for Indian and Non Indian Languages.

Key words: Hate Speech, Social Media, Machine Learning Algorithms, Indian and Foreign Languages.

I. INTRODUCTION

Now-a-days, the spread of online platforms has facilitated a global exchange of thoughts, views, and information. Though this has led to many positive interactions, it has also highlighted a negative aspect hate speech. Hate speech is when someone insults or downgrades others based on characteristics like race, religion, or gender. Hate speech not only erode societal harmony and individual happiness but also harm to democratic values. Hate speech uplift to violence, polarization and discrimination. The unchecked hate speech can cause serious real-world consequences. In 2018, throughout the Rohingya crisis in Myanmar, Facebook failed to stop hurtful posts. This contributed to spread violence against the Rohingya Muslim community [1].

In India, the National Crime Records Bureau (NCRB) reported that there were 323 registered hate speech cases, which increased to 1,804 cases in 2020. According to this report 500% rise reflected to online hate speech [2]. A Pew Research Center survey of U.S found 41% of adults Americans had experienced online harassment, with hate speech being a major factor [3]. An article published by the newspaper News Minute revealed that the internal Facebook documents had escalated violence in June 2020 during the Delhi Riots. It was reported that the inflammatory content on Facebook had increased by over 300% above the previous levels [4].

Hate speech can spoil society, individual well-being, and democratic values. As hate speech grows online, it's crucial for online platforms and governments to find ways to notice and prevent it. Finding hate speech manually is nearly impossible due to the huge amount of content posted daily; the researchers are solving this problem by using Machine Learning (ML) algorithms, especially focusing on Natural Language Processing (NLP). These algorithms analyze text patterns and meanings to spot hate speech. However, these algorithms don't work similarly well for all languages because of language differences, cultural differences, and the lack of proper training data. Research on hate speech detection focuses on foreign languages like, English, Arabic, German, Danish, Greek, Turkish, Indonesian, South African and Portuguese as well as Indian languages.

India is a multilingual country with 22 official languages. The linguistic diversity and huge amount of internet users facilitates the need of development of NLP system for Indian Languages. According to the Census 2011[5] the 22 main languages have a combined 1.17 billion speakers in India. The linguistic complexity is further augmented by code-mixing such as Tanglish, Hinglish, and Manglish, script differences and dialectal variations. For example, hateful idioms in Hindi may be very different from one in Bengali or Tamil. Additionally, many hate speech detection models are built using English-centric datasets, which results in low accuracy when applied to Indian language contexts. To undertake this complex problem, advanced machine learning (ML) and natural language processing (NLP) techniques are used. The following figure demonstrates working of automated Hate speech detection system.

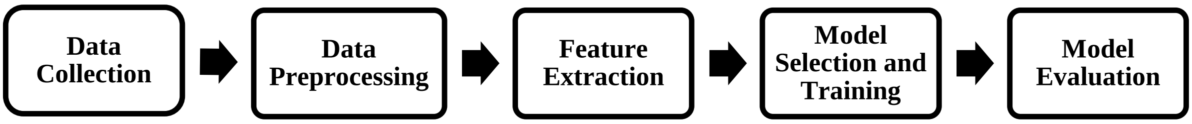

Figure 1: Hate Speech Detection Process

A methodical approach utilizing machine learning algorithms and natural language processing techniques are essential for detection of hate speech on social media. In the order to identify hate speech, text samples are first collected in data collection phase from social media platforms and categorized. The text content is edited by removed surplus characters, extra spaces, digits, special symbols, punctuations in data preprocessing phase. The meaning of the text is recovered through the techniques like bag-of-words, TF-IDF and word embeddings in feature extraction phase. Next, to detect and identify hate speech, a model is selected and trained by using deep learning and machine learning techniques as Logistic Regression, SVM, Random Forest, RNNs, LSTMs, or Transformers.

After splitting the data into training and testing sets, the models effectiveness is evaluated by using metrics like precision, recall, accuracy and F1-score.

II. TECHNIQUES USED FOR DEVELOPMENT OF HATE SPEECH DETECTION SYSTEM

Detection of hate speech is a motivating task due to the complexity and variability of language, cultural differences, and evolving nature of online content. A variety of techniques and methods have been developed to identify and combat hate speech online. Following are some of the commonly used techniques of hate speech detection.

A. Traditional Machine Learning Approach

Traditional machine learning techniques like, Naive Bayes. Random Forest, SVM and Logistic Regression are effective for hate speech detection. From literature survey, it is observed that Naive Bayes performed well with text features like TF-IDF, while SVM excels with TF-IDF or embeddings but needs balanced datasets. Logistic Regression is matched for linear features like n-grams, and Random Forest algorithms also worked efficiently [6].

B. Deep Learning Approach

Deep learning techniques improve hate speech identification by learning complex patterns. CNN (Convolution Neural Networks) detect spatial dealings in text, while RNN (Recurrent Neural Network) and LSTMs (Long Short-Term Memory networks) method deals with sequential data and capturing complex contexts. LSTM improve accuracy with character n-grams, TF-IDF, Bag-of-Word vectors and its mechanisms focus on boosting detection in longer texts and key sentence parts [6].

C. Transformer-Based Models Approach

Transformer-based models like RoBERTa {Robustly Optimized BERT), BERT (Bidirectional Encoder Representations from Transformers) and DistilBERT transform hate speech detection by capturing bidirectional perspective and understanding text complexity. BERT achieves state-of-the-art results, while RoBERTa and DistilBERT offer faster, efficient alternatives [7].

D. Hybrid Models Approach

Hybrid techniques enhance hate speech detection by combining multiple techniques.. Ensemble methods, such as combining Random Forest, SVM and CNN, influence on each model's strengths to achieve higher accuracy and robust performance [8].

III. RELATED WORK

Hate speech detection is a critical area of research, especially with the growing impact of online interactions by social media. The study on hate speech detection has been widespread and primarily focusing on global languages like English, German, and Spanish along with some European languages Surveys have provided important insights into existing challenges, methodologies, and issues for new researchers in analyzing their objectives. However, there is comparatively less research has been conducted in low resource languages including many Asian and Indian languages [9][10]. A major gap remains in multilingual hate speech detection, particularly in regional Indian and code-mixed languages. Initial research on hate speech detection has predominantly focused on English-language datasets due to its accessibility and ease of processing. A number of studies have broadened their scope to include diverse languages. For instance, the study [7] explored the rapid advancements in neural networks for detecting hate speech in code-mixed multilingual like Hindi-English, Tamil-English, Malayalam-English, Kannada-English datasets. In another study, researchers [11] investigated major Asian languages, particularly Indonesian, Arabic, Nepali to identify the most effective methods for hate speech detection and also examined how factors like vocabulary size, dataset quality and classification accuracy interrelate in this detection. Early studies on hate speech detection frequently used traditional machine learning algorithms like SVM, Decision Trees and Naïve Bayes. However, recent research has demonstrated that deep learning models, especially transformer-based models which enhance accuracy in detection. Some new perspectives have also emerged, such as the combined modeling of abusive and emotional language detection. In the study, researchers [12] took a novel approach to recognize the overlap between abusive language detection and sentiment analysis utilizing a Bengali dataset for their study. The researchers [13] conducted a survey on detecting hate speech in social networks using multiple languages' datasets, this study provides an overview of hate speech, correlated detection methods, focuses on the challenges and recommendations specific to Arabic hate speech detection. They also discusses outcome from existing multilingual hate speech datasets and concludes that deep learning techniques like RNNs and CNNs are commonly used for this task. The earliest survey on detecting hate speech in code-mixed text was conducted by [14]. The authors focused on a variety of methods for identifying hate speech in a Hindi-English code-mixed language. They also emphasized that hate speech can exist in code-mixed languages and highlight the challenges posed by such datasets with , informal grammar, spelling inconsistencies and make the detection more difficult. Deep Learning CNN, LSTM and BiLSTM performed well by using word embeddings like Word2Vec, FastText, and GloVe for both Indian and foreign languages [15].

Our study on hate speech detection across multiple languages is well-organized and systematic. The key contribution of this study is to provide a comparative analysis of existing research of different methodological approaches, in hate speech detection systems. It highlights the applications of machine learning and deep learning techniques in both foreign and Indian Languages and a comprehensive review of the results reported in various research studies. Table 1 and Table 2 gives the detailed review for foreign and Indian Languages.

Table 1: Review of Hate speech detection developed for foreign languages

Author and Year	Reff-rence No.	Language	Datasets Platform	Techniques Used	Feature Extraction Techniques	Performance Outcomes
W.Z.Wase-em and D. Hovy (2016)	[16]	English	Twitter	Grid Search.	Word N-grams	F1: 64.58%, Precision: 64.39%, Recall: 71.93%.
Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., &Tesconi, M. (2017)	[17]	Italian	Facebook	SVM, LSTM	POS, N-grams, Lemma sentiment polarity N-grams	F1: 71.8%-72.8%.
I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, (2017)	[18]	Indonesian	Twitter	Naive Bayes, SVM, Random Forest, Ensemble	Various	F1: 79.8% (Ensemble approach).
M. A. Fauzi and A. Yuniarti (2018)	[19]	Indonesian	Twitter	LSTM	Pre-trained Glove embeddings (300 dimensions), Count Vectorizer (CV)	Training F1: 0.78, Test F1: 0.72.
P. Fortuna, J. R. da Silva, J. Soler-Company, L. Wanner, and S.	[20]	Portuguese	Twitter	SVM, LSTM, GRU, SGD	TF-IDF, Unigrams	LSTM/GRU F1: 0.89, SVM F1: 0.80, SGD Recall: 0.61.

Nunes (2019)						
Z. Pitenis, M. Zampieri, and T. Ranasinghe (2020)	[21]	Greek	Twitter	SVM, Naïve Bayes, Logistic Regression, BERT Random Forest,	Doc2Vec, TF-IDF	F1: 0.90 (English), 0.56 (Danish), 0.67 (Greek).
K. A. and T. D (2020)	[22]	English, Danish, and Greek	Twitter	SVM, Random Forest, Gradient Boosting	Word N-grams, Character N-grams, Syntactic Features, TF-IDF	Accuracy: 0.671.
T. Mandla, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, and J. Schäfer (2020)	[23]	German, English and Hindi	Twitter	BERT, SVM, CNN, BiLSTM	TF-IDF	BERT F1: 0.9496.
C. Coltekin (2020)	[24]	Turkish	Twitter	Naive Bayes, SVM, Logistic Regression, Random Forest, Decision Tree	Word N-grams	F1: 93.5%.
Niraula, N. B., Dulal, S., & Koirala, D (2021)	[25]	Nepali	Facebook, Twitter, YouTube, Nepali Blogs, and News Portals.	LR, SVM, RF	nGrams	RF outperformed with F1: 0.86 (Non-Offensive), 0.73 (Offensive).
Vargas, F. A., Carvalho, I., de Góes, F. R., Benevenuto, F., & Pardo, T. A. S. (2021).	[26]	Brazilian Portegese	Instagram	NB, SVM, MLP and LR Evaluation.	Unigram, Tf-IDF	F1 Score:0.85
Markov, I., Gevers, I., & Daelemans, W. (2022, June)..	[27]	Dutch	Facebook, Twitter,	BERT RobBERT SVM Ensemble method	nGrams	F1 Score 75.3% SVM
Thapa, S., Rauniyar, K., Shiwakoti, S., Poudel, S., Naseem, U., & Nasim, M. (2023).	[28]	Nepali	Twitter	Naive Bayes, Decision Tree, XGBoost , AdaBoost, NepBERT,	TF-IDF	F1-score of 0.68
Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A & Silva, C. (2024).	[29]	Portugese	YouTube, Twitter	CNN LSTM BERT, BERTimbau,	Fastext, CBOW	F-score of 87.1%

From Table1 it is observed that variety of hate speech detection systems is available for non-Indian language like English. It is also observed that ML algorithms like NB, SVM, RF, LSTM, DT LR, CNN, and RNN are used for development of hate speech detection systems.

Table 2: Review of hate speech detection developed for Indian languages

Author	Ref-ernce No.	Language	Datasets Platform	Techniques Used	Feature Extraction Techniques	Performance Outcomes
Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018).	[30]	Hindi-English code mixed text	Twiiter	SVM and Random Forest algorithms	n-grams, character n-grams, and word2vec embeddings	SVM Accuracies 71.7%
Chakraborty, P., & Seddiqui, M. H. (2019).	[31]	Bengali	Facebook	Linear SVC, Naive Bayes RF, Gated Recurrent Unit (GRU})	nGrams, TFIDF	Random Forest Accuracy 52.20% GRU Accuracy 70.10% (increase about 18%0.
M. M. Khan, K. Shahzad, and M. K. Malik (2020)	[32]	Urdu	Twitter	LR, SVM, NB, CNN, RF	Count Vector, TFIDF	LR + Count Vector: F1 Score 0.756 (Offensive-Hate Speech), 0.906 (Neutral-Hostile).
B. R. Chakravarthi and team (2020)	[33]	Tamil and Malayalam	YouTube Comment	CNN, XLM-Roberta	Lexicon	F1 Score: 0.65 and 0.74.
Adeep Hande (2020)	[34]	Kannada Code Mixed	YouTube Posts	LR,Decision Tree, KNN, Naive Bayes, SVM	TFIDF	LR F1 Score 0.61
N. Romim, M. Ahmed, H. Talukder, and M. S. Islam (2021)	[35]	Bengali	YouTube Facebook	SVM, LSTM, BiLSTM	CBOW, Word2Vec, FastText, BengFast	Accuracy: 87.5% and 86.55%.
D. Saha and team (2021)	[36]	Tamil	YouTube	LR, RF, XLM-Roberta, CNN	TFIDF	Average weighted F1 Score: 0.77.
C. Vasantharajan and U. Thayasivam (2021)	[37]	Tamil, Kannda and Malayalam	YouTube	BERT	M-BERT	F1 Scores: 0.73, 0.70, and 0.96.
S.Jayanthi and A.Gupta (2021)	[38]	Kannada, Tamil, Malyalam	YouTube	mBERT, XLM-Roberta	mBERT	F1 Scores: 70.30, 76.94, and 96.76.
Bharathi B and Agnusimmaculate Silvia A. (2021)	[39]	code mixed Dravidian languages Malyalam, Tamil and Kannada	YouTube	BERT	TFIDF, Counter Vectorizer, BERT	F1 Scores: 0.95, 0.73, and 0.70.
Qinyu Que and team (2021)	[40]	Kannada	Social Media (not given in Paper)Twitter Facebook	XLM-Roberta, LSTM	XLM-Roberta, K-Fold Cross-Validation	F1 Score: 0.33.
Konthala Yasaswini (2021)	[41]	Malyalam, Tamil and Kannada code mixed Dravidian languages	Social Media	BERT, XLM-Roberta, CNN-BiLSTM, DistilBERT, ULMFIT, mBERT, ALBERT	Doc2Vec, GloVe	Kannada: 0.7277, Malayalam: 0.9603, Tamil: 0.7895 (F1 Scores).
Pathak V. Joshi M. Joshi P. Mundada M.Joshi T.	[42]	Malaya-lam, Tamil	Twitter, YouTube,	SVM, RF, LR Multinomial NB,	nGrams, TFIDF	Malyalam F1 Score: 0.77, Tamil F1

IJCRT

(2020)				Decision Tree		Score: 0.87.
Sudhir Kumar Mohapatra (2021)	[43]	Odia - English code mixed data	Facebook Public Pages Post	SVM, NB, RF	nGrams, TFIDF, Word2Vec	F1 Score: 73%.
P. Ram, and team (2021)	[44]	Tamil	YouTube Comments /Posts	LR, SVM	TFIDF	F1-Score improved by 2.3% (Logistic), 2.2% (SVM).
A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, (2021)	[45]	Hindi, Marathi	Twitter	CNN, LSTM, BERT, IndicBERT, RoBERTa	FastText	Marathi: CNN Precision: 0.910, LSTM Accuracy: 0.859, Recall: 0.847; Hindi: CNN F1: 0.760, Recall: 0.760.
D. Gajbhiye, S. Deshpande, P. Ghante, A. Kale, and D. Chaudhari (2021)	[46]	Marathi	Twitter	LR, RF	TFIDF	Accuracy: RF 77.70%, LR 75.95%.
V. Bansal, M. Tyagi, R. Sharma, V. Gupta, and Q. Xin (2022)	[47]	13 indic code mixed Languages	Comments posts	XLM-Roberta, mBERT, MuRILBERT, IndicBERT	BIGRU, Emoji2Vec Library	F1 Score: XLM-Roberta: 87.603.
P. K. Roy, S. Bhawal, and C. N. Subalalitha (2022)	[48]	Malyalam, Tamil	Twitter ,Youtube	LR, RF, SVM, BERT, MuRIL	Ensemble, Deep Learning	F1 Scores: 0.700 (Malayalam), 0.829 (Tamil).
A. K. R., P. Poornachandran, S. V. G., G. Rajendran, V. KS., V. Vijayan, and A. Ram,(2022)	[49]	Malayalam	Twitter	mBERT	mBERT	F1 Score: 0.85.
G. Swathi, A. Sharanya, M. Akhila, and B. Sra (2023)	[50]	Gujrathi	Twitter	SVM, RF	TFIDF, nGram	Accuracy: 95.6%.
N. Narayan, M. Biswal, P. Goyal, and A. Panigrahi (2023)	[51]	Bengali, Asamees , Bodo, Sinhala, Gujrathi	Twitter, Facebook,Youtube	XLM-R, BERT, LSTM, SVM	Emoji2Vec	F1 Scores: Bengali: 0.6707, Bodo: 0.83009, Sinhala: 0.83493. Assamese: 0.70525,
S. Devi, K. S., and A. K. Madasamy (2023)	[52]	Tamil, Code mixed	Twitter, And Helo app	Hierarchical Attention Network (HAN)	nGrams, Word2Vec	F1 Score: 0.88.
J. Boda (2023)	[53]	Gujrathi	Twitter, Facebook Instagram	Decision Tree, RF, SVM, LR, Gaussian NB, KNN	TFIDF, Bag of Words	F1 Score: 0.67.
] J. K. Mim, M. Oussalah, and A. Singhal (2023)	[54]	Asamees , Bengali and Bodo	Twitter, Facebook,YouTu-be	XLM-Roberta, IndicBERT, L3Cube, ChatGPT3	Self-Annotation	F1 Scores: Assamese: 72.2, Bengali: 73.4, Bodo: 76.2.
A. Joshi and R. Joshi (2023)	[55]	Gujrathi, Asamees , Bengali	Twitter,Facebook,Youtube HASOC2023	BERT, SBERT, GujrathiBERT, IndicBERT, BengaliBERT	-	F1 Scores: Gujrathi: 73.24, Bengali: 70.65, Tamil: 77.03.
K. Ghosh, A. Senapati, M. Narzary, and M. Brahma (2023)	[56]	Bodo, Asamees	Facebook,YouTube	NB, SVM, LSTM, BiLSTM, CNN	TFIDF, nGrams, Word2Vec, Emoji2Vec	F1: mBERT: 0.87 (Bodo).
P. P. Bansod (2023)	[57]	Hindi	Twitter,HAS	DistilBERT,	FastText,	MuRIL

			OC dataset	MuRILBERT, LR, Decision Tree	GloVe	Embeddings F1 Score: 0.73.
Farsi, S., Hoque, A., Hossain, E. J., Ahsan, S., Das, A., & Hoque, M. M. (2024).	[58]	Telgu Code Mixed	Social Media Comments GitHub Repository	LR, SVM, Ensemble, CNN, BiLSTM, Indic-SBERT, XLM-R	TFIDF, FastText, Word2Vec	F1 Scores: SVM: 0.65, IndicSBERT: 0.70.
S. Sangeetham, S. C. Vinay, K. Rajan G, A. Abishna, and B. Bharathi (2024)	[59]	Tamil	LT-EDI-EACL 2024	LR, SVM, MLP, RF, KNN, Decision Tree	TFIDF	F1 Score: SVM: 0.77, Accuracy: 0.75.

It is observed from above table that many ML & DL algorithms are used for the development of hate speech detection system for Indian language. However more focus is on Bengali & Tamil language. As far as Marathi is concern very few efforts are reported in the literature.

Early research 2016 - 2020 primarily relied on traditional ML models such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression, using TF-IDF and n-grams for feature extraction. From 2020-2022deep learning approaches, including CNN, LSTM, and BiLSTM, demonstrated improved performance using word embeddings such as Word2Vec, FastText, and GloVe. Recent studies 2022–2024 increasingly adopt transformer-based architectures like BERT, XLM-Roberta, IndicBERT, and MuRIL, achieving superior F1 scores, particularly in multilingual and code-mixed langauges.

IV. METHODOLOGY

The methodology of this study had a systematic approach to review and analyze the hate speech detection techniques in both foreign and Indian languages. The methodology used for this study was reviewed with the research papers published from 2016 to 2024. These research papers collected from peer reviewed journals, well-established databases, conferences, and reputable sources like IEEE Explore, ACM, Springer, and arXiv.

A review guided that, a number of datasets has been developed for hate speech detection in various languages. These datasets have given valuable contribution of research in low-resource and multilingual settings. The HASOC dataset used to find Hindi and Marathi labeled data [45]. The researchers used NEIHS dataset low resource Indian languages like Bodo and Assamese [56]. Another datasets, like KanCMD [34] for Kannda and Hate-Aler[36] for Tamil helped in detection of hate speech and offensive contents identification. In Foreign languages, SSN_NLP_MLRG dataset used for collection of English, Danish, Greek twitter comments [22]. The OGTD Dataset used to detect hate content in Greek teweets and and the HateBR dataset identified Brazilian Portuguese hatred instagram comments [21][26].

To ensure a systematic and convincing review, the evaluation considered on the source of data, hate speech detection techniques used, feature extraction methods, and performance metrics, including precision, recall, F1-score, and accuracy. The study of hate speech detection moved towards hybrid and advanced transformer-based models like BERT, IndicBERT, XLM-Roberta and MuRIL, which achieved higher accuracy for code-mixed texts as well as multilingual detection systems. The following Table III presents the comparative analysis of techniques used for hate speech detection.

Table 3: Comparative analysis of techniques used for Hate Speech Detection

Approach	Common Techniques Used	References	F1 Score Range
Traditional Machine Learning	Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Decision Tree (DT)	[16], [17], [18], [19], [21], [22], [23], [25], [26], [30], [31], [32], [34], [36], [37], [42], [43], [44], [46], [50], [53]	Foreign Languages:-0.60 - 0.80 Indian Languages:-0.65-0.80
Deep Learning	CNN, LSTM, BiLSTM, GRU, MLP	[17], [19], [20], [23], [25], [27], [29], [32], [35], [37], [38], [41], [44], [45], [46], [52], [56], [58]	Foreign Languages:- often greater than-0.70 Indian Languages:- often greater than-0.75
Transformer-Based Learning	BERT, XLM-R, mBERT, IndicBERT, RobBERT	[21], [23], [24], [27], [28], [33], [37], [38], [39], [40], [41], [47], [48], [49], [51], [54], [55],[56], [57], [58], [59]	Foreign Languages:- often greater than-0.85 Indian Languages:- often greater than-0.85

Hybrid Approaches	Ensemble methods (e.g., CNN-LSTM, SVM + BERT, RF + TF-IDF)	[18], [21], [25], [26], [27], [29], [33], [37], [41], [42], [48], [54], [57], [58]	Foreign Languages:- often greater than-0.85 Indian Languages:- often greater than-0.85

Table 4: Different platforms used for dataset development

Social Media Platform	**References**
Twitter	[16], [18], [19], [20], [21], [22], [23], [24], [28], [29], [30], [32], [42], [45], [46], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59]
Facebook	[17], [25], [27], [28], [31], [35], [43], [51], [52], [53], [54], [55]
YouTube	[25], [29], [33], [34], [35], [36], [37], [38], [39], [42], [44], [48], [51], [54]
Instagram	[26], [29], [53]
Blogs & News Portals	[25]

The social media platforms used for hate speech detection research summary from Table 4 indicates that Twitter is the most frequently used platform that appearing in the majority of studies due to its real-time, text-heavy nature. YouTube and Facebook are also widely used specially for multilingual and multimedia content. Instagram has limited research focus, likely due to its visual-centric format. Blogs and news portals are not used frequently but provide long-form text for analysis. The choice of platform influences dataset characteristics that Twitter favors short text-based analysis and YouTube/Facebook supports multimodal approaches.

The feature extraction techniques in hate speech detection research summarizes from Table 1 and 2, that n-Grams and TF-IDF are the commonly used methods, especially in traditional machine learning models like SVM and RF. Word embeddings (Word2Vec, FastText, CBOW, Doc2Vec) are widely applied in deep learning models such as LSTMs and CNNs for better contextual understanding. Emoji2Vec is emerging for detecting sentiment in social media texts. The choice of technique depends on the model, where as it is noted that machine learning favoring TF-IDF/n-Grams and deep learning benefiting from word embeddings.

V. CONCLUSION

From the literature review, it is evidence that, there is a limited development in automated detection of hate speech systems for several Indian languages, including Telugu, Kannada, Punjabi, Hindi, and Marathi. The research indicates that only a few techniques have been in use for the development of hate speech classifiers in Indian languages. These techniques include Support Vector Machine, Naive Bayes, and Decision Trees. The limited research and the restricted use of techniques highlight the need for more comprehensive and diverse research with incorporating deep learning and transformer-based models like BERT, IndicBERT, and XLM-Roberta in detection of hate speech across Indian languages. The future research should focus on integrating context-aware techniques to enhance detection accuracy.

REFERENCES

[1] Lee, R. (2019). Extreme speech| extreme speech in Myanmar: The role of state media in the Rohingya forced migration crisis. *International Journal of Communication*, *13*, 22.

[2] Sen, P. (2023). Hate speech & media laws in India: A critique. *International Journal of Law Management & Humanities, 6*(4). [ISSN 2581-5369].

[3] Pew Research Center. (2021, January 13). *The state of online harassment*. Pew Research Center. Retrieved from https://www.pewresearch.org.

[4] S`uri, S., & Randive, M. (2023). Social Media as a Platform for Instigating and Waging War.

[5] INDIA. 2011. Census of india, 2011. https: //www.censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf.

[6] Pen, H., Teo, N., & Wang, Z. (2024, June). Comparative Analysis of Hate Speech Detection: Traditional vs. Deep Learning Approaches. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 332-337). IEEE. 6].

[7] Dowlagar, S., & Mamidi, R. (2021, December). *A survey of recent neural network models on code-mixed Indian hate speech data*. Conference paper, LTRC, IIIT-Hyderabad.

[8] Ojo, O. E., Ta, T. H., Gelbukh, A., Calvo, H., Sidorov, G., & Adebanji, O. O. (2022). Automatic hate speech detection using deep Neural networks and word embedding. *Computación y Sistemas*, *26*(2), 1007-1013.].

[9] Alrehili A (2019) Automatic hate speech detection on social media: a brief survey. In: 2019 IEEE/ACS 16th international conference on computer systems and applications (AICCSA). IEEE, pp 1–6

[10] Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the 5th international workshop on natural language processing for socialmedia, pp 1–10

[11] Dhanya L, Balakrishnan K (2021) Hate speech detection in Asian languages: A Survey. In: 2021 International conference on communication, control and information sciences (ICCISc) 1:1–5 (IEEE)

[12] Rahman AI, Akhand Z-E, Noor MAU, Islam J, Mahtab M, Mehedi MHK, Rasel AA, et al (2022) Comparative analysis on joint modeling of emotion and abuse detection in Bangla language. In: International conference on advances in computing and data sciences. Springer, pp 199–209

[13] Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In 6th International Conference on Computer Science and Information Technology, Vol. 10.

[14] A Comparative Study of Different State-of-the-Art Hate Speech Detection Methods for Hindi-English Code-Mixed Data Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae Insight SFI ResearchCentre for Data Analytics, Data Science Institute, National University of Ireland Galway {priya.rani, shardul.suryawanshi, koustava.goswami, bharathi.raja, theodorus.fransen, john.mccrae}@insight-centre.org

[15] Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., Suieuova, N., & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *14*(5), 396. Retrieved from http://www.ijacsa.thesai.org

[16] W. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proceedings of the NAACL Student Research Workshop, pp. 88-93, 2016.

[17] Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. Italian Conference on Cybersecurity (ITASEC17), 86–95.

[18] Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (n.d.). *Hate speech detection in the Indonesian language: A dataset and preliminarystudy*. Machine Learning and Computer Vision Laboratory, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia.

Retrieved from ika.alfina@cs.ui.ac.id, rio.mulia@ui.ac.id, ivan@cs.ui.ac.id, yudo.ekanata51@ui.ac.id

[19] M. A. Fauzi and A. Yuniarti, "Ensemble Method for Indonesian Twitter Hate Speech Detection," 2018.

[20] P. Fortuna, J. R. da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A Hierarchically-Labeled Portuguese Hate Speech Dataset," INESC TEC, FEUP, University of Porto, Porto, Portugal and NLP Group, ETIC, Pompeu Fabra University, Barcelona, Spain.

[21] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive Language Identification in Greek," 2020.

[22] K. A. and T. D., "SSN_NLP_MLRG at SemEval-2020 Task 12: Offensive Language Identification in English, Danish, Greek using BERT and Machine Learning Approach," Department of CSE, SSN College of Engineering, India, email: kalaiwind@gmail.com, theni_d@ssn.edu.in.

[23] T. Mandla, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, and J. Schäfer, "Overview of the HASOC track At FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages," 2020.

[24] C. Coltekin, "A corpus of Turkish offensive language on social media," in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 6174-6184.

[25] Niraula, N. B., Dulal, S., & Koirala, D. (2021, August). Offensive language detection in Nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 67-75).

[26] Vargas, F. A., Carvalho, I., de Góes, F. R., Benevenuto, F., & Pardo, T. A. S. (2021). Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. *arXiv preprint arXiv:2103.14972*

[27] Markov, I., Gevers, I., & Daelemans, W. (2022, June). An ensemble approach for Dutch cross-domain hate speech detection. In *International conference on applications of natural language to* information *systems* (pp. 3-15). Cham: Springer International Publishing.

[28] Thapa, S., Rauniyar, K., Shiwakoti, S., Poudel, S., Naseem, U., & Nasim, M. (2023). Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023* (pp. 2346-2353). IOS Press.

[29] Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., ... & Silva, C. (2024). Leveraging transfer learning for hate\speech detection in portuguese social media posts. *IEEE Access*.

[30] Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media* (pp. 36-41).

[31] Ishmam, A. M., & Sharmin, S. (2019, December). Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)* (pp. 555-560). IEEE.

[32] M. M. Khan, K. Shahzad, and M. K. Malik, "Hate Speech Detection in Roman Urdu."

[33] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, and J. P. McCrae, "Overview of the track On Sentiment Analysis for Dravidian Languages in Code-Mixed Text."(2020)

[34] A. Hande, R. Priyadharshini, and B. R. Chakravarthi, "KanCMD: Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection," 1Affiliation, 2Affiliation, 3Affiliation.

[35] N. Romim, M. Ahmed, H. Talukder, and M. S. Islam, "Hate Speech detection in the Bengali language: A dataset and its baseline evaluation," in Shahjalal University of Science and Technology, Kumargaon, Sylhet 3114, Bangladesh.

[36] D. Saha, N. Paharia, D. Chakraborty, P. Saha, and A. Mukherjee, "Hate-Alert@DravidianLangTech-EACL2021: Ensembling Strategies for Transformer-based Offensive language Detection," Indian Institute of Technology, Kharagpur, India.

[37] C. Vasantharajan and U. Thayasivam, "Hypers@DravidianLangTech-EACL2021: Offensive language identification in Dravidian code-mixed YouTube Comments and Posts," Department of Computer Science and Engineering, University of Moratuwa, Colombo,SriLanka.

[38] S. M. Jayanthi and A. Gupta, "SJAJ@DravidianLangTech-EACL2021: Task-Adaptive Pre-Training of Multilingual BERT models forOffensive Language Identification," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021.

[39] B. B. and A. S. A., "SSNCSE NLP@DravidianLangTech-EACL2021: Offensive Language Identification on Multilingual Code Mixing Text," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association forComputational Linguistics, 2021.

[40] Q. Que, G. Wang, and S. Jia, "Simon @ DravidianLangTech-EACL2021: Detecting Offensive Content in Kannada Language," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021.

[41] K. Yasaswini, K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, and B. R. Chakravarthi, "IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021.

[42] Pathak, V., Joshi, M., Joshi, P., Mundada, M., & Joshi, T. (2021). Kbcnmujal@ hasoc-dravidian-codemix-fire2020: Using machine learning for detection of hate speech and offensive code-mixed social media text. *arXiv preprint arXiv:2102.09866*.

[43] S. K. Mohapatra, S. Prasad, D. K. Bebarta, T. K. Das, K. Srinivasan, and Y.-C. Hu, "Automatic Hate Speech Detection in English-Odia Code Mixed Social Media Data Using Machine Learning Techniques," 2021.

[44] P. Ram, M. T. Meeradevi, V. M. Vinod, G. A. Gothainayaki, A. S. Anusha, and T. Agalya, "Comparative Analysis for Offensive Language Identification of Tamil Text Using SVM and Logistic Classifier," Electronics and Communication Engineering, Kongu Engineering College, Erode, Tamil Nadu,

[45] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and Offensive Speech Detection in Hindi and Marathi," 2023.

[46] D. Gajbhiye, S. Deshpande, P. Ghante, A. Kale, and D. Chaudhari, "Machine Learning Models for Hate Speech Identification in Marathi Language," 2023.

[47] V. Bansal, M. Tyagi, R. Sharma, V. Gupta, and Q. Xin, "A Transformer Based Approach for Abuse Detection in Code Mixed IndicLanguages," Bharati Vidyapeeth’s College of Engineering, India; Institute of Computer Science, University of Tartu, Estonia; Jindal Global Business School, O.P. Jindal Global University, India; Faculty of Science and Technology, University of the Faroe Islands, Faroe Islands, 2022.

[48] P. K. Roy, S. Bhawal, and C. N. Subalalitha, "Hate speech and offensive language detection in Dravidian languages using deepensemble framework," Department of Computer Science & Engineering, Indian Institute of Information Technology, Surat, India; School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha, India; Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.

[49] A. K. R., P. Poornachandran, S. V. G., G. Rajendran, V. KS., V. Vijayan, and A. Ram, "MalHate: Hate Speech Detection in Malayalam Regional Language," Centre for Internet Studies and Artificial Intelligence, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam, India.(2022)

[50] G. Swathi, A. Sharanya, M. Akhila, and B. Sra, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in Gujarati Tweets," ZKG International, ISSN: 2366-1313.

[51] N. Narayan, M. Biswal, P. Goyal, and A. Panigrahi, "Hate Speech and Offensive Content Detection in Indo-Aryan Languages: A Battle of LSTM and Transformers," Z-AGI Labs, India.

[52] S. Devi, K. S., and A. K. Madasamy, "The Effect of Phrase Vector Embedding in Explainable Hierarchical Attention-based Tamil Code-Mixed Hate Speech and Intent Detection," 1Department, 2Department, (Member, IEEE) and (Senior Member, IEEE).2023

[53] J. Boda, "Investigating Hostile Post Detection in Gujarati: A Machine Learning Approach," Research Square, 2023.

[54] J. K. Mim, M. Oussalah, and A. Singhal, "Cross-Linguistic Offensive Language Detection: BERT-Based Analysis of Bengali, Assamese, & Bodo Conversational Hateful Content from Social Media," 2023.

[55] A. Joshi and R. Joshi, "Harnessing Pre-Trained Sentence Transformers for Offensive Language Detection in Indian Languages," 2023.

[56] K. Ghosh, A. Senapati, M. Narzary, and M. Brahma, "Hate Speech Detection in Low-Resource Bodo and Assamese Texts with ML-DL and BERT Models," 2023.

[57] P. P. Bansod, "Hate Speech Detection in Hindi," Master's Projects, 2023.

[58] Farsi, S., Hoque, A., Hossain, E. J., Ahsan, S., Das, A., & Hoque, M. M. (2024). Hate and offensive language detection in Telugu code-mixed text using sentence similarity BERT. *Proceedings of EACL 2024: DravidianLangTech*, CUET_Binary_Hackers.

[59] S. Sangeetham, S. C. Vinay, K. Rajan G, A. Abishna, and B. Bharathi, "Algorithm Alliance@LT-EDI-2024: Caste and Migration Hate Speech Detection," Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India.