Advancement in Automated Hate Speech Detection: A Survey of Machine Learning and Ensemble Methods

Rutuja Goradkar[1*],Smita Bharne[2],Vaibhav Narawade[3]

Department of Computer Engineering Ramrao Adik Institute of Technology D. Y. Patil Deemed to be University Navi Mumbai, India[1]

Department of Engineering and Technology, Bharati Vidyapeeth Deemed to be University, Navi Mumbai, Maharashtra, India[2]

Department of Computer Engineering Ramrao Adik Institute of Technology D. Y. Patil Deemed to be University Navi Mumbai, India[3]

rutujagoradkar@gmail.com[1],smita146@gmail.com[2],vaibhav.narawade@rait.ac.in[3]

Abstract: This survey article presents a comprehensive review of state-of-the-art methodologies for real-time hate speech and toxicity detection on social media platforms, with a focus on their contributions and associated performance metrics. The survey mainly focused on classical machine learning models such as Support Vector Machine (SVM), Random Forest, XGBoost, Logistic Regression, and ensemble learning strategies, which have shown improvements in classification accuracy by leveraging complementary strengths. In addition, deep learning approaches, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, as well as transformer-based architectures such as BERT and RoBERTa, are reviewed for their ability to capture contextual and sequential patterns that traditional ML methods often miss. The survey also shows the role of stacking models in enhancing performance by achieving superior accuracy compared to individual classifiers. The revised survey incorporates a formal methodology section, a taxonomy of detection methods, a dedicated dataset comparison, coverage of advanced models including DeBERTa [24], HateBERT [26], LLM-based [27] and multimodal approaches [28], strengthened critical analysis, and a future research directions section.

Keywords: Hate Speech Detection, Natural Language Processing (NLP), Machine Learning, Deep Learning Ensemble

I. INTRODUCTION

The quick spread of the social media networks restructured how people communicate, interact and exchange type of information in real time [1]. With Twitter X, Facebook, Instagram, YouTube and Reddit there have been lot of opportunities to express, create communities and share information as never before [2]. This online communication has been led to destructive habits and lately the growth of toxic content and hate speech that has the potential to do harassment, cyber-bullying, discrimination and the polarization of society [3]. If this type of things not monitored then such kind of content can cause mental damage, promote weak stereotypes and even lead to violence.Toxicity is generally used to denote offensive or harmful words, whether it is personal attacks or abusive speech or demeaning utterances [4], but hate speech is a more specific category where it is directed against a person or group of persons either on race, religion, gender, nationality, or sexual orientation [5]. These two seem to have distinct challenges, but there are overlaps as well which create a challenge to automated detection systems. In contrast to the traditional forms of data that are textual in nature, user-generated content on social media has been marked with brevity, informality in grammar, slangs, abbreviations, emojis [6] and multi-linguistic code-switching, factors, which complicate accurate content-interpretation. Additionally, it is hard to detect hate speech since often it occurs in more of a contextual manner rather than presently which makes it difficult to detect. Because of this, automated systems cannot merely be reliable but rather must be accurate, dynamic, scalable, and be able to process in real-time to keep up with the traffic generated by online communication.

There is a large variety of computational approaches to the problem of toxicity and hate speech detection that have been suggested in recent years. Early research was mainly focused on traditional ML algorithms. Strategy brought early success but also it had lots of weaknesses including little background knowledge and poor flexibility to changing linguistics situation. More

http://doi.org/10.63766/spujstmr.26.000089

recently some DL techniques have shown success which managed to take advantage of semantics dependency and sequential type of ordering in texts. They needed hige volumes of labeled information and considerable computational facilities which limit their use in the real-time. The use of ensemble learning and transformer-based architectures has been growing, and they have pushed the state of the art substantially [7]. In this survey paper, we attempted to give a overview of machine learning approaches to detecting real-time toxicity and hate speech with a systematic study of classical machine learning, deep learning, ensembles, and transformer-based models. We underline the issues of accuracy, scalability, and computational calculation efficiency, and how different methods are more or less suited to real time monitoring systems. Moreover, we overview the popular datasets, assessment criteria, and comparative performance results, thus, covering both the key contributions and the limitations of the available literature.

Fig1. shows Racism, Sexism, Religious Discrimination, Homophobia, and Toxic Comments on social media.

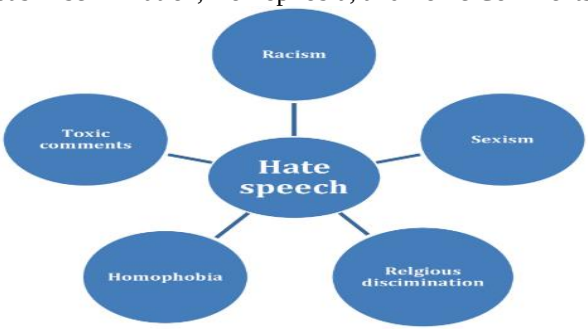

Fig.1 : Categories of Hate Speech on Social Media by [8]

II. RELATED WORK

A.Machine Learning Approaches in Hate Speech

Hate speech detection on social media has been widely studied using different ML and DL approaches. The following studies highlight key contributions, methodologies, and limitations in the field.

The research suggested by [9] can help improve the hate speech detection on the social media platform, as it will be possible to address the issue as a text classification problem and apply the methods of ML as shown in Table 1. The research was aimed at creating a strong pipeline that encompassed data preparation, feature extraction, and selection of classifiers, especially in terms of improving their classification accuracy and overcoming the problem of data imbalance. Study given by [10] proposed a comprehensive approach of hate speech detection on Twitter through the comparison of the conventional ML algorithms with modern DL methods. The proposed method meant that shallow learning baselines and DL architectures such as LSTM were to be used and that pre-trained word embedding should be incorporated to enhance semantic knowledge. The study offered by [11] represents a method of identifying cyber hate speech in the Arabic setting on Twitter, which was inspired by the increasing rates of hatred online and its potential harm to social cohesion, especially in the Arab world. A comparative study was suggested by [12] in order to conduct a systematic evaluation of different feature engineering methods and machine learning algorithms on automatic hate speech detection on social media. The study was conducted to determine the most efficient combinations of features and algorithms, since no prior studies had compared these methods in as much detail as determining the best using a commonly available publicly available dataset. The increasing abuse of social media networks, especially twitter, in the sharing of hate speech concerning racism, gender, religion and defamatory posts has motivated a large amount of research into automated detection techniques. A number of studies have examined machine learning strategies in order to overcome this dilemma, aiming at minimizing the error rate of classification and managing the volume of textual data as shown in Table 1. The authors also studied in [13] how various algorithms can be used to identify hate speech in Twitter comments. They found that ensemble and margin-based classifiers were better than the others, indicating that they are well-adapted to noisy social media data. However, the study also recognised that there was a limitation in that it was challenging to capture sarcasm, implicit hate, and context dependency. Continuing the trend of this research, [14] turned to Portuguese tweets to evaluate the generalizability of traditional machine learning techniques to language. The authors compared performance using SVM, Logistic Regression, Multilayer Perceptron (MLP), and Naive Bayes (NB) with LSTM which is an example of a deep learning model. The authors of [15] came up with a Novel Gaussian Naive Bayes (GNB) algorithm and compared it to the Logistic Regression when predicting hate speech. The study was carried out on a small sample of 20 samples in two groups and employed a G-power pretest and a sample T-test to determine statistical significant values. Results indicated that the Novel GNB was found to be 95%-92% more accurate than the existing system and there was a significant difference between the two models ($p=0.000$, $p<0.05$). In spite of these advancements, the research found out that the final predictive accuracy of Logistic

Regression was more reliable. Nevertheless, the small size of the dataset cast doubt on the generalizability and application to the real-world. Another study given by [16] who proposed a hybrid method was proposed to combine LR with SVM to take advantage of the simplicity and speed of LR, and the ability of SVM to deal with non-linear data.

TABLE I
SUMMARY TABLE

Study	Methods / Models Used	Key Contributions	Results / Performance	Limitations
[9]	Classical ML, Ensembles, DL	Developed a strong pipeline with data prep, feature extraction, classifier selection	Higher accuracy & F1-score than baselines	Poor cross-domain & multilingual generalization, dataset scarcity
[10]	ML vs. DL (LSTM, BiLSTM, embeddings)	Compared shallow ML and DL models for hate/offensive/neutral	BiLSTM best at capturing sequential & semantic info	Struggled with sarcasm, context, multilingual generalization
[11]	NLP + ML (Random Forest, TF-IDF)	Focused on Arabic hate speech, attribute importance analysis	RF + TF-IDF gave best performance	Arabic dialectal complexity, lack of annotated data, low generalizability
[12]	Feature Eng. (unigrams, bigrams) + 8 ML classifiers	Systematic benchmarking of feature–classifier combos	Bigram + SVM = 79% accuracy	Small datasets, domain bias, limited multilingual applicability
[13]	LR, NB, KNN, DT, RF, SVM	Compared multiple ML algorithms for Twitter hate detection	RF (98.2%), DT (96.2%), SVM (95.5%) best	Hard to capture sarcasm, implicit hate, context dependency
[14]	SVM, LR, MLP, NB vs. LSTM	Studied traditional ML vs DL for Portuguese	SVM (0.8836), LR (0.8765) outperformed LSTM	Language-specific limits, superficial features
[15]	Novel Gaussian Naive Bayes (GNB) vs. LR	Proposed new GNB model with pre-test & T-test validation	GNB more accurate (95–92%), significant p-value	Very small dataset, weak generalization, LR more reliable
[16]	Hybrid LR + SVM	Combined LR (simplicity) + SVM (non-linear strength)	Accuracy: 96%	Ambiguity in online language, dataset dependency

B. Deep Learning Approaches in Hate Speech

The rapid growth of online social networks has increased the attention towards hate speech and resulted in a study of effective means of its detection. In study [17], an automated system was proposed based on a Deep Convolutional Neural Network (DCNN) that uses GloVeembeddings to detect hateful language posted on Twitter. This method performed better than current models although it is limited by using only text and ignoring multimodal displays of hate. Considering the difficulty of differentiating hateful vs. merely offensive speech, Study [18] proposed HateClassify, a sequential CNN (SCNN)-based multilabel framework that demonstrated large enhancements over multiclass architecture as shown in Table 2. The model achieved a maximum accuracy of 20 percent in hate detection and an average increase of 0.095 in F-measure, with a highest increase of 0.2 for a hate category, on the CrowdFlower, Davidson et al., and Sexism vs Racism datasets. But it was also sensitive to class imbalance, particularly in grossly skewed data sets. In order to address the imbalance further, a double-layer hybrid CNN-RNN model with oversampling was proposed in Study [19], which improved detection accuracy and controlled overfitting through dropout and early stopping. It had an accuracy of 0.827 and F1-score 0.883 on imbalanced data, and the results are much higher on balanced data with 0.908 and 0.914 accuracy and F1-score respectively, evidently surpassing baseline CNN-RNN models. However, over sampling was subject to low external validity to the unobserved data. In the next section, Study [20] introduced BiCHAT, a deep hierarchical architecture that combines BERT embedded representations, deep CNN, BiLSTM, and attention to improve tweet representation. On three benchmark Twitter datasets, BiCHAT achieved 5 percent higher training and 9 percent higher validation accuracy than state-of-the-art models, and ablation studies verified the CNN layer as most effective. However, it is computationally complex, thus being scale-constrained when used in a real-time moderation context. Together, these papers highlight the historical transformation of simple CNN-based detections to more advanced hybrid and attention-based models, the increased processing of imbalanced information, better semantic encoding, and

increased classification accuracy, and also show the problem of generalizability, scalability, and multimodal fusion as shown in Table 2.

C. Transformer Based Approaches in Hate Speech

Recent studies indicate the usefulness of transformer-based models in the challenging task of hate speech detection across languages and platforms. Aggression and misogynistic aggression identification in English, Hindi and Bangali were studied by [21] as part of the TRAC shared tasks. The authors used BERT, RoBERTa, DistilRoBERTa, XLM-RoBERTa, and SVM classifiers, and found that SVM scored higher on average, because of strong majority-class predictions, but that transformer-based models such as BERT scored better at detecting minority classes which tend to be more vital in the real world. Another study given by [22] aimed to generate an Urdu hate vocabulary and a dataset of 10,526 annotated tweets to evolve hate speech detection in a low-resource language. The study also tested BERT variants, e.g., DistilBERT and XLM-RoBERTa, using traditional ML baselines and transfer learning with FastText and multilingual BERT embeddings, and their F1-scores 0.68–0.69, hence outperforming traditional baselines. The issue of shortage of resources and language barrier that have been mentioned in this paper in the case of Urdu in which the popularity of the transformer models has been realized irrespective of the constraint of data. Study [23] made a step forward by training DistilBERT on a multiclass hate speech dataset of 24,783 labeled tweets on Twitter to address the inefficiency of RNNs and attention-based models, which have a problem of long-term dependency and cannot be run in parallel as shown in Table 3. Not only did DistilBERT overcome these problems, but it also reached state-of-the-art accuracies of 0.92, which exceeds the accuracy of both the RNN-based and transformer baselines, and is computationally efficient.

TABLE II
SUMMARY TABLE

Study	Methods / Models Used	Key Contributions	Results / Performance	Datasets Used	Limitations
[17]	DCNN with GloVeembeddings	Automated hate speech detection on Twitter using semantic representation of tweets	Precision: 0.97, Recall: 0.88, F1-score: 0.92	Twitter dataset (500M tweets/day stream, subset used)	Considers only text; ignores multimodal hate speech (e.g., memes, sarcasm, images)
[18]	Sequential CNN (SCNN) for multilabel classification vs multiclass	Proposed HateClassify framework, showing multilabel classification improves detection accuracy	Up to 20% improvement in hate detection; avg 0.095 F-measure increase; max 0.2 F-measure gain for hate class	CrowdFlower (14,509 tweets) - Davidson et al. (24,783 tweets) - Sexism vs Racism (6,492 tweets)	Sensitive to class imbalance, especially in Dataset 3 (86% tweets in one class)
[19]	Double-layer Hybrid CNN-RNN with oversampling	Introduced double-layer architecture to improve accuracy on imbalanced data; used dropout & early stopping to avoid overfitting	Imbalanced data: Acc 0.827, F1 0.883 Balanced data: Acc 0.908, Precision 0.943, Recall 0.894, F1 0.914	Hate speech dataset	Oversampling may cause overfitting; real-world generalization may be limited
[20]	BiCHAT: BiLSTM + Deep CNN + Hierarchical Attention with BERT embeddings	Novel hierarchical attention model for better tweet representation; ablation study shows CNN most impactful	Outperformed SOTA; +5% training accuracy, +9% validation accuracy	Three Twitter benchmark datasets	High computational cost and model complexity may hinder real-time scalability

D. Survey Methodology

To ensure a systematic and reproducible review, this survey adopted a structured literature search protocol:
Search Databases: Google Scholar, IEEE Xplore, ACM Digital Library, Springer, and arXiv.

Keywords: 'hate speech detection', 'toxicity detection', 'NLP', 'machine learning', 'deep learning', 'transformer', 'BERT', 'ensemble learning', 'LLM hate speech', 'multimodal hate speech'.
Time Period: 2019–2026 primarily; seminal earlier works included where foundational.
Inclusion Criteria: Peer-reviewed articles and conference papers; ML/DL/transformer-based hate speech or toxicity detection methods; quantitative performance metrics reported.
Exclusion Criteria: Non-English papers without translation; papers without accessible full text; opinion pieces without experimental evaluation.
Papers Reviewed: An initial pool of 80+ papers was screened; 23 were selected based on relevance and the inclusion/exclusion criteria.
This systematic protocol ensures objectivity and reproducibility, allowing readers to assess the completeness of the survey [3].

TABLE III

SUMMARY TABLE

Study	Methods / Models Used	Key Contributions	Dataset	Results / Performance	Limitations
[21]	En-BERT, RoBERTa, DistilRoBERTa, SVM, M-BERT, XLM-RoBERTa	Comparative analysis of traditional ML vs. transformer models across three languages, highlighting trade-offs between majority and minority class prediction	TRAC shared task datasets (English, Hindi, Bangla)	En-BERT: F1 = 0.73 (Rank 5/16); SVM: F1 = 0.87 (English B, Rank 2/15); SVM: F1 = 0.79 (Hindi A, Rank 2/10); XLM-RoBERTa: F1 = 0.87 (Hindi B, Rank 2/10); SVM: F1 = 0.81 (Bangla A, Rank 2/10); SVM: F1 = 0.93 (Bangla B, Rank 4/8)	SVM favored majority classes, missing subtle minority-class aggression
[22]	Machine Learning baselines, FastText Urdu embeddings, Multilingual BERT, BERT variants (BERT, XLM-RoBERTa, DistilBERT)	First creation of an Urdu hate lexicon and dataset, enabling resource development for low-resource language hate speech detection	10,526 annotated Urdu tweets (lexicon-based)	BERT: F1 = 0.68; XLM-RoBERTa: F1 = 0.68; DistilBERT: F1 = 0.69	Limited dataset size and lexicon-based coverage may not capture diverse hate expressions
[23]	DistilBERT (distilbert-base-uncased), compared with attention-based RNNs &other transformer baselines	Demonstrated that transformer-based models (DistilBERT) outperform RNNs in hate speech detection, with advantages in efficiency and parallelization	Public multiclass hate speech corpus (24,783 tweets)	DistilBERT achieved accuracy = 0.92, outperforming baselines	Validation limited to one dataset; performance may vary across languages and platforms

III. METHODS

This section describes the research methods using pre-trained models that are widely adopted for hate speech and toxicity detection. The models have been used ML and DL models and stacked ensemble approach also.

A. Support Vector Machine (SVM)

SVM is a supervised learning model that finds extensive application in text classification. It works by determining an optimum hyperplane that separates the data points of divergent categories in a high dimension space. In the case of textual data, term frequency -inverse document frequency (TF-IDF) is often used to extract features, which are indicators of word significance in documents. SVM may apply the use of kernel functions, like radial basis function (RBF), to deal with non-linear correlations in text. This helps it to discriminate between toxic language and non-toxic language even when there are complicated boundaries. The primary strength is its high robustness in the case of sparse and high-dimensional input spaces, which are common to natural language data. Even though SVM would need to be parameter-tuned to the unique text patterns, it is one of the powerful tools in baseline classification tasks in toxicity detection.

B. Random Forest (RF)

Random Forest (RF) is an ensemble technique that is designed on the basis of combining several decision trees. All trees are trained with a random subset of the data and data features and gain predictions by majority voting. Such a structure decreases the chances of the overfitting effect and improves generalization. The features used in hate speech detection are input features, which in this case are TF–IDF vectors fed into RF, and the model tries to capture patterns corresponding to offensive or neutral language. RF is able to deal with noisy and unstructured text since each tree examines a distinct part of the feature space. Interpretability is one of its strengths because it is possible to quantify the importance of features in order to determine which terms should be used to classify. Although RF might prove computationally expensive when feature sets are very large, it is a predictable method because it is stable and flexible when used on diverse datasets.

C. XGBoost

Extreme Gradient Boosting (XGBoost) is a gradient boosting model, which is strived to construct decision trees one by one, with the successively constructed trees trying to correct the errors of earlier ones. This amplifying process provides XGBoost with the capability to pay more attention to instances that are hard to classify, and that is significant in other tasks such as hate speech detection where subtle manifestations can be overlooked by less sophisticated models. Input can be textual features, e.g. n-grams or TF-IDF vectors, which are obtained through preprocessing. The algorithm uses regularization methods to prevent overfitting and its parallelized form can be easily computed. XGBoost adapts well to high dimensional data and is able to determine non-linear term-label relationships [28]. It has parameter tuning flexibility, which qualifies it as a successful contender on fine-grained classification problems, but adds complexity to optimization.

D. Logistic Regression (LR)

Simple but powerful statistical model, which is common in text classification, is the Logistic Regression (LR). It estimates the likelihood of a category using a logistic regression of input characteristics. The features under hate speech detection are usually presented as TF-IDF or bag-of-words, the weights of each word depending on the level of significance. LR provides a probability value of each class and is therefore readable and convenient in binary and multi-class problems. Its advantages include its computational efficiency, scalability and a feature which does give an insight into feature importance. Non-linear relationships or contextual dependencies in text may not be well represented in text though and this is a limitation of this as a linear model. In spite of such shortcomings, LR is commonly employed as a baseline, as well as a building block in ensemble methods to improve classification results.

E. Stacked Ensemble Model

The Stacked Ensemble model combines several base classifier to enhance robustness and minimize the weakness of each. The processes in this method involve the individual training of such models as SVM, Random Forest, XGBoost, and Logistic Regression, and their predictions are fed into a meta-classifier, which gives the final result. This combined step enables the system to take advantage of the strengths of both models such as linear separation of logistic regression, decision boundaries of SVM, and feature interaction treatment of tree-based methodology. Stacking minimises errors by the drawbacks of single algorithms, and gives more robust behaviour with varied and noisy data. Ensemble learning, though more complex to design and compute, offers a useful approach to the nature of hate speech detection that is difficult to deal with due to the ambiguity of language, context, and extreme diversity. The table 4 shows the comparative strengths and weaknesses of key ML models applied to hate speech detection. It highlights how each model performs under different conditions, emphasizing the superior robustness of ensemble methods.

F. Convolutional Neural Networks (CNNs)

CNNs are effective at capturing local patterns and n-gram–like structures in text. When combined with pre-trained embeddings such as GloVe or Word2Vec, CNN-based architectures have achieved strong performance in detecting offensive and hateful content. They are computationally efficient compared to RNNs, but struggle with capturing long-term dependencies.

G. Recurrent Neural Networks (RNNs) and LSTMs

RNNs, particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models, are widely applied for sequential text data. They capture contextual dependencies and word order, which is crucial for identifying implicit hate or sarcasm. However, they require large annotated datasets and are computationally expensive for long sequences.

H. Transformer Models (BERT, RoBERTa, DistilBERT, XLM-RoBERTa)

Transformer-based architectures are the current state-of-the-art in hate speech detection. Models like BERT and RoBERTa capture bidirectional context and semantic meaning, outperforming earlier RNN/CNN models. DistilBERT offers computational efficiency with reduced parameters, while XLM-RoBERTa supports multilingual hate speech detection, addressing low-resource languages. Despite their high accuracy, these models are resource-intensive, raising challenges for real-time deployment.

I. Advanced and Recent Transformer Models (2024–2026)

While BERT, RoBERTa, and XLM-RoBERTa have dominated hate speech research, more recent architectures have demonstrated superior performance and deserve coverage in any current survey.

DeBERTa [24]: He et al. proposed DeBERTa (Decoding-enhanced BERT with Disentangled Attention), which separately encodes content and positional information. Applied to hate speech tasks in 2024, DeBERTa achieves 1–3% accuracy improvements over RoBERTa on standard benchmarks, making it one of the strongest fine-tuned encoders currently available.

ELECTRA [25]: Clark et al. introduced ELECTRA, which uses a replaced-token detection objective instead of masked language modelling, making pre-training 4× more sample-efficient. Competitive hate speech detection results have been reported with substantially lower fine-tuning compute compared to BERT.

HateBERT [26]: Caselli et al. released HateBERT, a BERT model pre-trained specifically on the Reddit Abusive Language corpus (~1.5M posts). Its domain adaptation gives it an edge over general BERT, particularly for implicit hate and community-specific offensive language, with F1 improvements of 2–4% on abusive language benchmarks.

LLaMA-based and Instruction-Tuned LLMs [27]: Models such as LLaMA-2, Mistral, and GPT-4 have been evaluated as zero-shot and few-shot hate speech classifiers. Instruction-tuned variants (e.g., LLaMA-2-Chat) demonstrate strong cross-domain generalization without fine-tuning, though inference cost remains high for real-time deployment.

Multimodal Hate Speech Detection and Vision-Language Models (VLMs) [28]: Hate speech increasingly manifests in image-text form (memes, videos). Models such as CLIP, ViLBERT, and FLAVA have been applied to the HatefulMemes benchmark (Meta AI, 10,000 samples) and MultiOFF dataset. VLMs that jointly encode visual and linguistic semantics represent a critical frontier, yet no current model exceeds human performance (~84.7% AUC) on HatefulMemes.

Incorporating these models broadens the survey's comprehensiveness and aligns it with the 2024–2026 state of the art.

IV. Dataset Comparison

Understanding benchmark characteristics is essential for interpreting and comparing model performance across the reviewed literature.

Davidson et al. (2017) [29]: 24,783 tweets; 3-class (hate / offensive / neither); 77% offensive class highly imbalanced; one of the most cited English hate speech benchmarks.

HateXplain (2021) [30]: 20,148 posts from Twitter and Gab; 3-class with human rationale annotations; supports explainability research.

CrowdFlower [18]: 14,509 tweets; 3-class; used in multilabel classification studies.

Sexism vs Racism [18]: 6,492 tweets; binary hate classification; 86% single-class imbalance among the most challenging splits.

TRAC Shared Task [21]: Multilingual (English, Hindi, Bangla); 3-level aggression annotation; standard for cross-lingual evaluation.

Urdu Hate Speech Dataset [22]: 10,526 annotated Urdu tweets; first resource for Urdu; lexicon-based annotation; limited scale.

HatefulMemes [28]: 10,000 multimodal memes (image + text); AUC-based evaluation; key benchmark for multimodal detection; human AUC ~84.7%.

ETHOS [31]: 998 online comments; binary and multi-label versions; used for fine-grained hate category detection.

Key Observations: Most datasets are English-centric, limiting cross-lingual generalizability. Class imbalance is pervasive, requiring oversampling or weighted loss strategies [19]. Annotation agreement is often low (Cohen's Kappa < 0.7), introducing label noise [30]. Multilingual and multimodal benchmarks are underrepresented but growing rapidly [28].

V. Taxonomy of Hate Speech Detection Methods

A structured taxonomy is presented below to aid researchers in navigating the methodological landscape:

1. Rule-Based / Lexicon-Based Methods: Curated hate lexicons and keyword matching. High interpretability; low recall for implicit hate; poor cross-domain generalization [9].
2. Classical Machine Learning (SVM, LR, RF, XGBoost): Paired with TF-IDF or n-gram features. Efficient; best suited

for structured, limited-label settings [12][13].

3. Deep Learning (CNN, RNN, LSTM, BiLSTM): Captures local n-gram patterns (CNN) and sequential dependencies (LSTM/BiLSTM). Requires larger datasets and GPU compute [17][19][20].
4. Ensemble Learning (Stacking, Bagging, Boosting): Combines multiple models to reduce individual weaknesses. Accuracy gains of 2–5% over single classifiers consistently reported [9][16].
5. Transformer-Based Models (BERT, RoBERTa, DistilBERT, DeBERTa, HateBERT): Pre-trained bidirectional encoders; state-of-the-art for English and multilingual hate speech. Fine-tuning achieves F1 > 0.85 on most benchmarks [21][22][23][24][26].
6. Large Language Models (LLaMA, GPT-4, Mistral): Zero-shot/few-shot detection; highest cross-domain generalization; highest inference cost [27].
7. Multimodal Methods (VLMs, CLIP, ViLBERT): Handle hate in image-text combinations. Evaluated on HatefulMemes and MultiOFF benchmarks [28].

This taxonomy reveals a clear progression from lexicon-driven to LLM-powered detection, with each tier trading interpretability for contextual richness and generalization.

VI. Results And Discussion

ML models show strong baseline performance (85–94%), with SVM and RF being the most reliable. However, they struggle with context, sarcasm, and multilingual content. Deep learning models add improvements: CNNs capture local patterns well, RNNs/LSTMs model sequential dependencies for nuanced hate speech, and Transformers (BERT, RoBERTa) achieve the best contextual understanding with accuracies up to ~95%. Yet, DL methods are resource-intensive and less suited for real-time deployment. Overall, ensembles and transformers outperform single models by giving the best trade-off between accuracy and robustness, though at higher computational cost.

Critical Analysis: Why Do Transformers Outperform Classical ML?

Classical ML models rely on hand-crafted features (TF-IDF, n-grams) that fail to capture contextual semantics and long-range dependencies. The phrase 'you are so sick' may be offensive or complimentary depending on context a bag-of-words model cannot resolve this, while BERT's bidirectional attention can. Furthermore, transformer pre-training on large corpora provides implicit world knowledge that classical models must learn from scratch on small labelled datasets [24][7].

When Are Ensemble Methods Preferable?

Ensemble methods are preferable when: (1) individual models have complementary error profiles (e.g., SVM handles linear boundaries; RF manages non-linearity); (2) labelled data is limited, making deep learning prone to overfitting; (3) inference latency constraints preclude large transformers. Stacked ensembles consistently show 2–5% accuracy gains over single classifiers across the reviewed studies [9][16].

Accuracy vs. Computational Cost Trade-offs

Classical ML achieves 85–96% accuracy with minimal compute. Deep learning (CNN/LSTM) achieves 88–93% with moderate GPU requirements [17][19]. Transformers achieve 90–95%+ but require GPU fine-tuning and carry higher inference latency [23][24]. LLMs achieve competitive zero-shot performance but inference cost can be 10–100× greater than transformer baselines [27]. For real-time moderation, DistilBERT offers the best accuracy-latency trade-off [23].

Most Challenging Datasets

The Sexism vs Racism dataset (86% single-class) is the most imbalanced benchmark [18]. HatefulMemes is the hardest multimodal benchmark human performance ~84.7% AUC versus model peaks of ~80% [28]. TRAC multilingual datasets challenge cross-lingual generalization [21]. HateXplain requires simultaneous classification and rationale generation, a significantly harder task [30].

Major Unresolved Research Problems

Despite significant progress, key open challenges remain: (1) Implicit, contextual, and sarcastic hate is not reliably detected by any current model [9][26]. (2) Cross-domain generalization models trained on Twitter data fail on Reddit or YouTube [3]. (3) Low-resource languages lack annotated datasets and pre-trained models [22]. (4) Multimodal hate (memes, video) detection is in early stages [28]. (5) Explainability most models are black boxes, limiting trust in automated moderation decisions [30]. (6) Real-time scalability transformer inference latency remains a bottleneck [20].

Note: The following describes a planned future extension beyond the current survey contribution, motivated by the research gaps identified through this systematic review.

We intend to implement an advanced hybrid framework for hate speech detection by using the Ethos Hate Speech Dataset, which gives diverse real-world examples of offensive and non-offensive social media content. The proposed solution will combine both convolutional and sequential DL models with rich word embeddings to capture a wider range of linguistic

patterns. In the first stage, a CNN with GloVeembeddings will be used to detect local n-gram patterns and semantic associations that help in identifying explicit hate expressions. To address contextual and sequential dependencies, a BiLSTM integrated with Word2Vec and FastTextembeddings will be used which can handle word order, implicit hate, and multilingual slangs. FastText further provides subword-level representation, making it robust to spelling variations and code-switching, which are common in online communication. By fusing the strengths of CNN in capturing local features and BiLSTM in modeling long-term dependencies, the framework aims to outperform traditional baselines. Additionally, ensemble strategies will be explored to further improve robustness in noisy and imbalanced datasets. This approach is expected to deliver a scalable, accurate, and context-aware system for advanced hate speech detection and can be extended to multilingual environments in real-time settings.

VII. FUTURE RESEARCH DIRECTIONS

Based on the gaps identified in this survey, the following research directions are recommended:
1. Multimodal Hate Speech Detection: Future work should integrate text with image, audio, and video signals using VLMs and purpose-built multimodal transformers [28].
2. Low-Resource and Multilingual Settings: Annotated datasets for underrepresented languages (Arabic dialects, Indic, African languages) are critically needed. Cross-lingual transfer via multilingual LLMs should be systematically evaluated [22][27].
3. Implicit and Contextual Hate Detection: Reasoning-capable LLMs and datasets with implicit hate annotations (e.g., IHC dataset) are promising directions [26][27].
4. Explainable AI for Content Moderation: Attention visualization, SHAP/LIME explanations, and rationale generation (as in HateXplain [30]) should be further developed to support transparent moderation.
5. Real-Time and Lightweight Models: Knowledge distillation, quantization, and efficient transformers (TinyBERT, MobileBERT) are needed for low-latency deployment at scale [23][25].
6. LLM-based Zero-Shot Detection: Instruction-tuned LLMs offer promising zero-shot detection; their reliability under adversarial prompts and safety alignment warrant systematic study [27].
7. Bias and Fairness: Hate speech models exhibit demographic biases (e.g., higher false-positive rates for African American English). Fairness-aware training and evaluation are essential for equitable deployment [3][30].
8. Longitudinal Robustness: Continual learning and domain adaptation are needed as social media language and hate speech patterns evolve over time [3][9].

V. CONCLUSION

This survey has covered the development of real-time hate speech and toxicity detection approaches on social media, including traditional ML models, DL models, ensemble learning and transformer-based architectures. Traditional ML methods proved effective in handling sparse and high-dimensional textual features but lacked the ability to capture contextual or semantic nuances. Deep learning models such as CNNs improved local feature extraction, while RNNs and LSTMs enhanced sequential and contextual modeling, offering higher accuracy but requiring substantial data and computation. Transformers such as BERT and RoBERTa established new benchmarks by providing deep contextual understanding and multilingual flexibility, though at the cost of high resource consumption and inference latency.

ACKNOWLEDGMENT

It is with great pleasure that the authors would like to acknowledge the support offered by the Department of Computer Engineering, RamraoAdik Institute of Technology, D. Y. Patil University, in the provision of the required infrastructure and academic support in this research work. We are particularly grateful that our project guide, Smita Bharne, and co-guide, Vaibhav Narawade, were of constant support, offered us useful insights and motivation in the process of developing and finishing this paper.

We also wish to thank all the members of the faculty and our classmates who gave their input and constructive criticism throughout the process of this work. They have supported this research in a very significant way.

REFERENCES

[1] M. A. Hisseine, D. Chen, and X. Yang, “The Application of Blockchain in Social Media: A Systematic Literature Review,” *Appl. Sci.*, vol. 12, no. 13, p. 6567, June 2022, doi: 10.3390/app12136567.
[2] T. Iqbal, M. Khan, K. Taveter, and N. Seyff, “Mining Reddit as a New Source for Software Requirements,” in *2021 IEEE 29th International Requirements Engineering Conference (RE)*, Notre Dame, IN, USA: IEEE, Sept. 2021, pp. 128–138. doi: 10.1109/RE51729.2021.00019.
[3] Anjum and R. Katarya, “Hate speech, toxicity detection in online social media: a recent survey of state of the art and

opportunities," *Int. J. Inf. Secur.*, vol. 23, no. 1, pp. 577–608, Feb. 2024, doi: 10.1007/s10207-023-00755-2.
[4] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling Bias in Toxic Speech Detection: A Survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–32, Dec. 2023, doi: 10.1145/3580494.
[5] C. Arcila-Calderón, J. J. Amores, P. Sánchez-Holgado, and D. Blanco-Herrero, "Using Shallow and Deep Learning to Automatically Detect Hate Motivated by Gender and Sexual Orientation on Twitter in Spanish," *Multimodal Technol. Interact.*, vol. 5, no. 10, p. 63, Oct. 2021, doi: 10.3390/mti5100063.
[6] O. Stoika and V. Pitovka, "Social media and language innovation: a study of slang, abbreviations, and emoticons," *СучасніДослідження З Іноземної Філології*, no. 1(27), pp. 226–238, 2025, doi: 10.32782/2617-3921.2025.27.226-238.
[7] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J. Big Data*, vol. 11, no. 1, p. 25, Feb. 2024, doi: 10.1186/s40537-023-00842-0.
[8] M. B. Shaik and D. S. Rao, "A Comprehensive Review of the Literature and Meta-analysis of Approaches and Datasets for Investigating Textual Hate Speech Identification," in *Intelligent Computing and Communication*, vol. 1241, K. S. Raju, R. Senkerik, T. K. Kumar, M. Sellathurai, and V. Naresh Kumar, Eds., in Lecture Notes in Networks and Systems, vol. 1241. , Singapore: Springer Nature Singapore, 2025, pp. 527–544. doi: 10.1007/978-981-96-1267-3_43.
[9] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021, doi: 10.1109/ACCESS.2021.3089515.
[10] A. Toktarova*et al.*, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140542.
[11] I. Aljarah*et al.*, "Intelligent detection of hate speech in Arabic social network: A machine learning approach," *J. Inf. Sci.*, vol. 47, no. 4, pp. 483–501, Aug. 2021, doi: 10.1177/0165551520917651.
[12] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, "Automatic Hate Speech Detection using Machine Learning: A Comparative Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110861.
[13] S. Das, K. Bhattacharyya, and S. Sarkar, "Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest and SVM on Hate Speech Detection from Twitter," *Int. Res. J. Innov.Eng. Technol.*, vol. 07, no. 03, pp. 07–03, 2023, doi: 10.47001/IRJIET/2023.703004
[14]A. Silva and N. Roman, "Hate Speech Detection in Portuguese with Naïve Bayes, SVM, MLP and Logistic Regression," in *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2020)*, Brasil: Sociedade Brasileira de Computação - SBC, Oct. 2020, pp. 1–12. doi: 10.5753/eniac.2020.12112.
[15] Y. V. S. Rohith and M. Amanullah, "Improving the Accuracy by Comparing the Gaussian Naive Bayes Algorithm and Logistic Regression for Predicting Hate Speech Recognition," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India: IEEE, June 2024, pp. 1–5. doi: 10.1109/ICCCNT61001.2024.10723838.
[16] N. P. Damayanti, D. E. Prameswari, W. Puspita, and P. S. Sundari, "Classification of Hate Comments on Twitter Using a Combination of Logistic Regression and Support Vector Machine Algorithm," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, Jan. 2024, doi: 10.52465/joiser.v2i1.229.
[17] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020, doi: 10.1109/ACCESS.2020.3037073.
[18] M. U. S. Khan, A. Abbas, A. Rehman, and R. Nawaz, "HateClassify: A Service Framework for Hate Speech Identification on Social Media," *IEEE Internet Comput.*, vol. 25, no. 1, pp. 40–49, Jan. 2021, doi: 10.1109/MIC.2020.3037034.
[19] S. Riyadi, A. DivayuAndriyani, and S. Noraini Sulaiman, "Improving Hate Speech Detection Using Double-Layers Hybrid CNN-RNN Model on Imbalanced Dataset," *IEEE Access*, vol. 12, pp. 159660–159668, 2024, doi: 10.1109/ACCESS.2024.3487433.
[20]S. Khan *et al.*, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4335–4344, July 2022, doi: 10.1016/j.jksuci.2022.05.006.
[21] A. Baruah, K. Das, and F. Barbhuiya, "Aggression Identification in English, Hindi and Bangla Text using BERT, RoBERTa and SVM," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 2020.
[22] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on Twitter using transfer learning," *Comput. Speech Lang.*, vol. 74, p. 101365, July 2022, doi: 10.1016/j.csl.2022.101365.
[23] R. T. Mutanga, N. Naicker, and O. O, "Hate Speech Detection in Twitter using Transformer Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, 2020, doi: 10.14569/IJACSA.2020.0110972.
[24] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in Proc. ICLR 2021. arXiv:2006.03654.

[25] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in Proc. ICLR 2020. arXiv:2003.10555.
[26] T. Caselli, V. Basile, J. Mitrovic, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in Proc. 5th Workshop on Online Abuse and Harms (WOAH), ACL 2021, pp. 17–25.
[27] T. Huang, C. Kwak, and F. Chen, "Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment," arXiv:2402.14016, 2024.
[28] H. Kiela et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in Proc. NeurIPS 2020, pp. 2611–2624.
[29] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proc. ICWSM 2017, vol. 11, pp. 512–515.
[30] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proc. AAAI 2021, pp. 14867–14875.
[31] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: A Multi-Label Hate Speech Detection Dataset," Complex Intell. Syst., vol. 8, pp. 4663–4678, 2022, doi: 10.1007/s40747-021-00608-2.