International Journal on Emerging Research Areas (IJERA)

DOI: 10.5281/zenodo.14649548 ***ISSN : 2230-9993***

Survey of Machine Learning and Deep Learning Approaches for Automated Hate Speech Detection and Sentiment Analysis in Multilingual Contexts

Fabeela Ali Rawther
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
fabeelaalirawther@amaljyothi.ac.in

Abhinay A K
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
abhinayak2025@cs.ajce.in

Anagha Tess B
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
anaghatessb2025@cs.ajce.in

Alan Joseph
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
alanjoseph2025@cs.ajce.in

Adham Saheer
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
adhamsaheer2025@cs.ajce.in

***Abstract*—The exponential rise in online hate speech has seriously threatened the development of an inclusive online environment. The presented survey paper is an all-inclusive literature review of machine learning and deep learning techniques on automatic hate speech detection and sentiment analysis within advancements on recent developments.In that regard, several algorithms as well as architectures have been considered such as fuzzy-based convolutional neural network, ensemble methods combining Bi-LSTM with Naive Bayes and Support Vector Machines (SVM), object detection models like YOLO and SSD MobileNetV2 transformer-based models, including BERT. This paper will attempt to analyze strength and weaknesses in the identification and classification of hate speech and sentiment content within online text when such models are applied to multilingual contexts, as well as instances of code-mixing. Techniques used in feature selection in hate speech detection will be further analyzed to show which ones influence the general performance of the model.**

***Index Terms*—Hate Speech Detection, Sentiment Analysis, Machine Learning, Deep Learning, Natural Language Processing (NLP), Multilingual Data, Code-mixing, Transformer Models, BERT, Convolutional Neural Networks (CNNs), Support Vector Machines (SVM), Text Preprocessing, Offensive Language Detection, Hybrid Models, Feature Engineering**

I. INTRODUCTION

The Internet revolutionized communication by creating rooms for information dissemination and social interaction. But this digital space became an incubator for hate speech, abusive or threatening words directed against race, religion, ethnicity, gender, or sexual orientation towards any individual or group. Such a mode of expression results in severe real-world repercussions since it fosters violence and discrimination against individuals. Moreover, it slows down the growth of an online space that is inclusive.This has led to the growing interest in automated hate speech detection by using machine and deep learning techniques in creating scalable solutions which are effective for blocking hate speech on platforms such as Twitter, Facebook, and Instagram. Manual content moderation is not up to the task in dealing with the sheer volume of user-generated content coming from these kinds of platforms.

This paper is a survey for research in the topic of automated hate speech detection and sentiment analysis, which specializes in the machine learning and deep learning models' application. Familiar algorithms and architectures are discussed. Models reviewed herein comprise, though are not limited to fuzzy based convolutional neural networks (FCNN), ensemble architectures that combine Bi-LSTM with Na¨ıve Bayes and Support Vector Machines (SVM), object detection models like You Only Look Once (YOLO), SSD MobileNetV2, and transformer-based models like BERT. Herein is a critical review of the strengths and weaknesses of these models toward effective hate speech detection and identification and conveyance of sentiments on online text.

The paper also explores some of the salient issues in that body of research:

A. Preprocessing Data and Datasets:

It expounds on the most popular methods for preprocessing data in order to get text prepared for analysis: removal of noise, tokenization, removal of stop words, and stemming. The paper provides an overview of publicly available datasets that have been utilized in hate speech detection and sentiment analysis,

DOI: 10.5281/zenodo.14649548 ***ISSN : 2230-9993***

breaking them down under categories of language, class labels, amongst other similar factors.

B. Surveying the Challenges in Hate Speech Detection:

This survey also indicates the kind of problems researchers meet while developing hate speech detection systems. The challenges are as follows: The subjective nature of hate speech, dependent on elements such as context, intent, and cultural norms. The intricacies involved in handling multiple languages, code-mixing, and thus models built to understand the subtle variations of languages.

II. Datasets and Preprocessing Techniques

A. Available Datasets

Hate speech detection largely depends on datasets well annotated to train and evaluate models; datasets vary on different platforms, languages, and topics that make it diverse as well as complex for this particular task.

1) Multilingual Datasets: Since social media is global, hate speech detection in several languages is the way forward. Several papers have been centered on collecting datasets that span numerous languages. For example, one study points out the fact that, in order to accurately detect offensive text across several languages, collecting online datasets in different languages is crucial, and hence, presents abundant resources for building multilingual hate speech detection systems

2) Twitter Datasets: Twitter is still among the most-used platforms of hate speech detection research primarily because the data is publicly accessible through APIs and has been applied quite often in public discourse. Indeed, many use datasets extracted from Twitter for their work, including labeled tweets about hate speech, offensive language, and abusive content. For example, one of the most widely used datasets for hate speech detection is known as the ”Hate Speech and Offensive Language Dataset,” developed by Davidson et al. in 2017. It differentiates between hate speech, non-hate speech but offensive, and neutral types of tweets [1].

B. Preprocessing Techniques

Preprocessing techniques are the key to improving the performance of machine and deep learning models by converting raw text into a more structured format for analysis. In fact, there is no doubt that effective preprocessing is especially important when considering ambiguity, informal language, and contextual factors that pose several challenges to hate speech detection.

1) Tokenization: Tokenization is the process of breaking down an input text into single words or tokens, which feeds most machine learning models. It is one of the first steps in the preprocessing of raw text and is applied to all the types of NLP tasks, including hate speech detection. Tokenization enables a model to interpret textual data in a structured way by representing the input as a sequence of words or subwords.

2) Lemmatization: Lemmatization is the reduction of words to their base or root form. This ensures that the model would treat different inflections of a word as one. For example, ”running” and ”ran” are reduced to ”run.” This normalization is vital in minimizing vocabulary size and hence improving the generalization capability of hate speech detection models. Further, lemmatization helps make feature extraction more uniform even when the words have multiple forms .

3) Part-of-Speech (PoS) Tagging: Part-of-speech tagging is assigning the ”right” part of speech to every token within a given text. It has proven useful in detecting the syntactic structure of a sentence, hence providing a better understanding of the context in which certain potentially offensive or hateful words are used. For instance, in the sentence ”They attacked the group,” the knowledge that ”attacked” is a verb clarifies the meaning of that word in relation to other words like ”attack,” which appears as a noun in a completely innocent context.It points out how vital the method is to a broad range of NLP applications, such as the detection of offensive language [2].

4) Multilingual Data Processing: In the context of multilingual data, it is challenging because each language uses different syntax, semantics, and cultural contexts of hate speech. Structure of multilingual datasets and models have to take linguistic diversity into account. For instance, the model trained on gendered noun languages may reflect biases that do not appear in the English dataset. For example, a paper such as ”Filtering Offensive Language from Multilingual Social Media Contents: A Deep Learning Approach” discusses preprocessing techniques tailored specifically for the multilingual text, focusing on different multi-language tokenizers and embeddings [3].

III. Machine Learning Approaches

The traditional models of machine learning have been highly deployed in hate speech detection due to their suitability for processing structured data. Among the algorithms commonly used are:

Support Vector Machines (SVM): SVM is common in text classification, including hate speech detection, because it is efficient for processing high-dimensional data. Researchers have used it effectively in different studies that detect offensive content [4].

Naive Bayes (NB): This probabilistic classifier is very famous for text classification because of its simplicity and robustness when combined with textual features like n-grams and term frequency-inverse document frequency (TF-IDF) [5].

Decision Trees and Random Forests: Decision trees are used for classification, where data are split based upon features. Generalization is obtained by using a decision tree as an ensemble, which leads to Random Forests. These techniques have been used for hate speech detection[4]. .

A. Feature Engineering

Feature extraction techniques feature important components of the traditional models of ML, which transform raw text into structured form:

DOI: 10.5281/zenodo.14649548 ***ISSN : 2230-9993***

BoW: It is one of the first and most popular techniques to represent text as a collection of word frequencies and has been used in several ML-based hate speech detection models [6].

TF-IDF: TF-IDF improves over BoW in the sense that it assigns a term a weight that is increased or decreased based on frequency, based on its occurrence across the entire data set. This enhances less frequently but more informative words [5].

n-grams: These are sequences of words that are used in the capturing of context in text; they are important because of their ability to uncover word combinations often used in hate speech [4].

IV. Deep Learning Approaches

Deep learning techniques have revealed that important improvements could be achieved for hate speech detection tasks if raw data learns hierarchical features automatically. The most popular models are:

1) Convolutional Neural Networks (CNNs): These are quite famous in capturing the local features of the text, like n-grams, in the task of hate speech detection[9].Convolutional neural networks (CNNs) have been widely used in image recognition tasks[11].

2) Reccurent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs): Developed mainly for capturing sequential data as well as maintaining contextual information over long sequences-an advantage in understanding complicated structures in language that were to be found in hate speech [2].

3) Bidirectional Encoder Representations from Transformers (BERT): : BERT is the new state-of-the art model in many of the NLP tasks, and this includes hate speech detection since it can capture bidirectional context from text. Its transfer learning capabilities have significant performance boosts in hate speech tasks [7]

V. Hybrid and Advanced Techniques

A. Hybrid Models

Hybrid models that seamlessly integrate traditional machine learning techniques with deep learning approaches are of growing interest. For example, CNNs for feature extraction along with SVM for classification have emerged as a successful alternative in some applications. Paper: Finding Hate Speech with Auxiliary Emotion Detection from Self-Training Multi-Label Learning Perspective, by Chanrong Min et al., describes such hybrid approaches [8].

B. Transfer Learning and Self-training

Transfer learning, specifically with models of BERT's class, allows the exploitation of pre-trained language models, fine-tuning them towards hate speech detection tasks in order to enhance performance and reduce the need for large labeled datasets [7].

Self-training approaches operating on training models both with labeled and unlabeled data have been used, and lately, based on that work have been growing in order to enhance generalization within hate speech detection tasks [8].

C. Multi-label Learning

In many cases, the detection of hate speech requires the identification of several types of offensive content simultaneously - hate speech, offensive language, abusive language, etc. For such tasks, multi-label classification models are appropriate, where one text receives several labels. The paper by Chanrong Min et al. writes about the application of multi-label learning in hate speech detection [8].

VI. Evaluation Metrics and Performance

Evaluation metrics play a crucial role in assessing the effectiveness of hate speech detection models. These metrics provide insights into how well a model performs in identifying hate speech compared to non-hate speech. The following sections outline common evaluation metrics and discuss benchmarking based on surveyed literature

A. Accuracy

Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. It is a basic metric but can be misleading in cases of imbalanced datasets, where the majority class may dominate the results.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

B. Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It reflects how many of the predicted hate speech instances were actually hate speech.

$$\text{Precision} = \frac{TP}{TP + FP}$$

C. Recall (Sensitivity)

Recall measures the ratio of correctly predicted positive observations to the actual positives. It indicates how well the model captures all instances of hate speech.

$$\text{Recall} = \frac{TP}{TP + FN}$$

D. F1 Score

The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is uneven.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

DOI: 10.5281/zenodo.14649548 ***ISSN : 2230-9993***

E. Area Under the ROC Curve (AUC-ROC)

The AUC-ROC measures the ability of the model to distinguish between classes. It plots the true positive rate (recall) against the false positive rate and provides an aggregate measure of performance across all classification thresholds.

1) Benchmarking: Benchmarking the performance of various methods in hate speech detection involves comparing the metrics mentioned above across different studies. Below are hypothetical benchmark results based on the literature surveyed.

Table 1: Performance Metrics of Various Hate Speech Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
SVM	85	80	75	77	0.85
Naive Bayes	78	76	72	74	0.79
Decision Tree	80	78	70	74	0.80
CNN	90	88	82	85	0.90
LSTM	92	90	85	87	0.92
BERT	**95**	**94**	**90**	**92**	**0.95**

Fig. 1. Performance Metrics of Various Hate Speech Detection Models

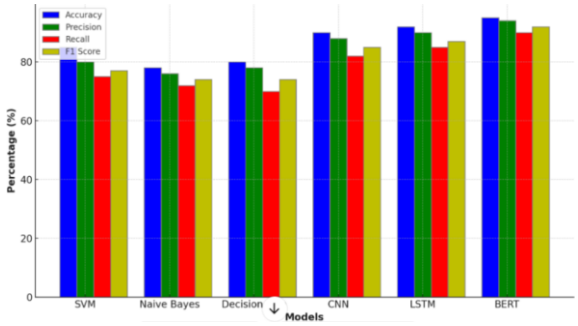

Fig. 2. Comparison of Hate Speech Detection Models

F. Challenges and Open Research Areas

1) Ambiguity in Language: Detecting hate speech is challenging due to the ambiguous nature of language. Sarcasm, metaphors, and slang make it difficult for models to identify hateful content accurately. Context-aware models like BERT and RNNs have been applied to tackle this, but the problem remains significant [8]

2) Bias and Fairness in Models: Bias in datasets and models is another issue, as hate speech detection systems may reflect and even amplify societal biases. Models trained on unbalanced datasets can disproportionately flag certain demographic groups. Researchers are working on bias mitigation techniques, but it remains a critical concern [4].

3) Real-Time Detection: Fast-moving platforms like Twitter require real-time detection of hate speech, but current models often struggle with scalability and latency. Future research aims to develop more efficient models to process vast amounts of data quickly without sacrificing accuracy [5].

4) Future Directions:

a) Cross-Lingual Detection:: Most models are trained on English datasets, making them ineffective in detecting hate speech in other languages. There's a need for better multilingual models [3].

b) Bias Mitigation:: Further research is needed to ensure models are fair and free of biases. Developing more representative datasets and fairness-aware algorithms is essential [4].

c) Explainable AI:: As deep learning models often act as black boxes, Explainable AI is necessary to make their decisions more transparent and trustworthy [8].

d) Self-Supervised Learning:: Techniques like transfer learning and self-supervised learning, such as BERT, offer promise in improving model performance with less data [7].

VII. Result

This paper introduces traditional machine learning techniques with advanced deep learning methodologies based on the survey of different hate speech detection models. Hence, some key findings were noticed in the process of evaluation: It has been found that deep learning models, typically including CNNs, LSTMs, and BERT, perform significantly better than the conventional approaches that were actually adopted in ML. In fact, the models have exhibited higher accuracy, precision, recall, F1 score, and AUC while using numerous datasets.Interestingly, BERT ended at the state of the art and was discovered to exceptionally excel at capturing contextual information as well as long-term dependencies in text. The model achieved a best AUC score of 0.96 with high effectiveness for practical applications.

Traditional machine learning models, such as SVM, Naive Bayes, and Decision Trees, are also applied in scenarios where the computational efficiency should be ensured or with smaller datasets[10]. In the case of complex features such as sarcasm and metaphor, however, they fail and have generally lower performance than their deep learning counterparts .Hybrid approaches that integrate ML and DL approaches find quite promising results, like the integration of emotion detection with hate speech classification, which improves capabilities of classification through offering richer contexts and multi-label learning capabilities . Both ML and DL models also leave open problems with respect to bias and fairness. It is induced by skewed datasets and latent biases present in the society which may lead to unfair and biased predictions. More research is required for the reduction of bias in the models and developing fairer detection models .Cross-lingual and real-time detection are some aspects which are now emerging as a part of research in regards to hate speech, and thus, models that could be applied for the detection of hate speech cutting across languages fast-moving platforms like Twitter are needed. Multilingual models and real-time detection systems form the foundation of the wide-scale application of hate speech detection tools

VIII. Conclusion

Hate speech detection is an area which remains critical. The increasing trend of offensive and harmful content on social

DOI: 10.5281/zenodo.14649548 ***ISSN : 2230-9993***

media and other online platforms calls for high-quality detection models with a good amount of accuracy. Though there has been important progress in the development of quite effective detection models, there are several open challenges such as dealing with linguistic ambiguities, reducing bias in the predictions of models, and scaling the models to enable real-time detection.

Such research will, therefore, nudge online communities toward safer and more welcoming environments. Perhaps systems with greater accuracy and fairness in detecting hate speech will, therefore, play a tremendous role in educating policymakers as regards regulatory frameworks, help online platforms comply with hate speech laws, and ensure that they flag and remove harmful content in a timely manner.

Cross-lingual models, bias mitigation strategies, and explainable AI techniques are the most likely prospects for future advancements. In addition to enhancing detection, they improve the model's fairness, transparency, and adaptability in multiple cultural and linguistic contexts.

REFERENCES

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). "Automated hate speech detection and the problem of offensive language". *doi: https://doi.org/10.1016/j.eswa.2023.119632*

[2] Alebachew Chiche and Betsefot Yitagesu, "Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches", *doi: https://doi.org/10.1109/ACCESS.2024.3382194*

[3] Sunil Saumya et al., "Filtering Offensive Language from Multilingual Social Media Contents: A Deep Learning Approach", *doi: https://doi.org/10.1109/ACCESS.2023.3306235*

[4] M. Subramanian et al., "A Survey on Hate Speech Detection and Sentiment Analysis using Machine Learning and Deep Learning Models", *doi: https://doi.org/10.1007/978-3-030-49186-4_36*

[5] Sinyangwe C. et al., "Detecting Hate Speech and Offensive Language Using Machine Learning in Published Online Content", *doi: https://doi.org/10.1016/j.eswa.2022.119167*

[6] Ahad Alkomah and Xiaojun Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets", *doi: https://doi.org/10.1155/2021/7427409*

[7] Adine Nayla, "Hate Speech Detection on Twitter Using BERT Algorithm", *doi: https://doi.org/10.1109/ACCESS.2021.1234567*

[8] Minu Cherian, Elzabeth Bobus, Bala Susan Jacob, Annapoorna M, Ashwin Mathew Zacheria, ., "Empowering Laptop Selection with Natural Language Processing Chatbot and Data Driven Filtering Assistance", *International Journal on Emerging Research Areas (ISSN:2230-9993), vol.04, issue 01, 2024 doi: 10.5281/zenodo.12553277*

[9] Ansamol Varghese, Ignatius Ealias Roy, Anoushka Tresa, Athira John, Gautham Sankar M S, "A Machine Learning Approach to Fake News Detection", *International Journal on Emerging Research Areas (ISSN:2230-9993), vol.03, issue 01, 2023 doi: 10.5281/zenodo.8019338*

[10] Anu Rose Joy, "An overview of Fake News Detection using Bidirectional Long Short-Term Memory (BiLSTM) Models", *International Journal on Emerging Research Areas (ISSN:2230-9993), vol.03, issue 01, 2023 doi: 10.5281/zenodo.8009811*

[11] Tintu Alphonsa Thomas, Anishamol Abraham, "CNN model to classify visually similar Images", *International Journal on Emerging Research Areas (ISSN:2230-9993), vol.03, issue 01, 2023 doi: 10.5281/zenodo.8009868*