Sigma J Eng Nat Sci, Vol. 44, No. 4, pp. 2313–2327, August, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.000169

Research Article

A comprehensive survey on hostile post detection techniques in social media: Challenges and future directions

Jagruti BODA[1,*], Keyur RANA[1]

[1]*Gujarat Technological University, 382424 India*

ARTICLE INFO

Article history
Received: 23 December 2024
Revised: 13 February 2025
Accepted: 22 July 2025

Keywords:
Dataset; Deep Learning; Hostile Post; Machine learning; Text Classification; Transformer Based

ABSTRACT

Hostile Post detection on social media is a crucial task that aims to identify harmful content such as hate speech, cyberbullying, fake news and offensive language. This survey provides a detailed survey of existing and emerging techniques for this task, including machine learning techniques, deep learning techniques, and hybrid techniques that integrate multiple approaches to improve model performance. We also explore the use of reinforcement learning, genetic algorithms, and pseudo-labeling techniques for optimising performance. There are several challenges identified, including dataset limitations, annotation inconsistencies, model interpretability, and ethical issues such as fairness and cultural sensitivity. The review also examines future trends in multilingual hostile post detection, cross-platform adaptability, and the increasing use of transformer-based architectures. Finally, we emphasize the importance of developing scalable, explainable, and context-aware solutions, and we outline future directions for research. This work is intended to serve as a reference for scholars and practitioners working toward safer online environments.

Cite this article as: Boda J, Rana K. A comprehensive survey on hostile post detection techniques in social media: Challenges and future directions. Sigma J Eng Nat Sci 2026;44(4):2313–2327.

INTRODUCTION

Social media has become an essential part of global communication, allowing users to engage, express and connect in real time. But the increasing usage of these platforms has also resulted in the development of hostile and destructive content such as hate speech, cyberbullying, foul language and misinformation. Such information may exacerbate social tensions, affect people's mental health and, in extreme situations, lead to real-world violence [1, 2]. The objective for this survey is the growing necessity to carefully review and compare numerous hostile post detection strategies used in both academic and operational contexts. The surge of online aggressiveness, particularly during politically sensitive moments and worldwide crises, makes the moderation of such content a serious task, especially in multilingual and low-resource settings. Hostile post detection is the automatic identification of damaging internet content that violates cultural and ethical norms. This effort is challenged by factors such as linguistic diversity, cultural differences in understanding anger, sarcasm, implicit meaning and developing communication styles

***Corresponding author.**
*E-mail address: jagrutiboda10@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

[3]. Manual moderation doesn't scale—you need AI-driven solutions that work at the size of major platforms. This survey uses a structured literature review process, focusing on hostile content detection strategies published from 2015 to 2024. The reviewed methods are classified into four categories, i.e., traditional machine learning, deep learning, hybrid models, and state-of-the-art methods. The study also discusses benchmark datasets, feature extraction methods, types of models, evaluation parameters and real-world application scenarios. To facilitate understanding, comparative tables and summarizations are also provided to summarize the main findings. Early systems used traditional machine learning approaches such as Support Vector Machines (SVM), Naive Bayes, and Random Forest [4] with customized features such as Bag-of-Words (BoW) and TF-IDF. However, these techniques fail to capture contextual or semantic information. Hence, there has been a rise of deep learning models [5] such as CNNs, RNNs, LSTMs and transformer-based models such as BERT [6]. These approaches get better semantic understanding and perform well in multilingual environments. Also, hybrid methods which combine deep learning with reinforcement learning, genetic algorithms, or optimization techniques have been proposed [7]. These are targeted at tackling difficulties such as data asymmetry, hyperparameter tuning and model adaptation. However, there is still effort to be made with regard to concerns such limited availability of labeled data, variability in annotations and the need for interpretable and fair models. This analysis is especially timely considering recent global events such the 2023–2024 Israel–Palestine conflict and India's national elections, where current automated systems failed to keep up with the scale and complexity of online hate. To this end, new ways are investigated by using advanced analytical frameworks like fractional calculus and fuzzy decision systems [8–11] to improve detection efficiency. The structure of this paper is as follows. In section 2, we define and categorize several categories of hostile content. Section 3 provides an overview of existing datasets. Section 4 discusses feature extraction methods from traditional to contextual embeddings. Section 5 presents the machine learning methods, and Section 6 concentrates on the deep learning methods. Section 7 presents hybrid and advanced strategies that combine multiple methodologies. Section 8 summarizes the comparative performance of ML, DL and hybrid approaches. Section 9 describes the most used evaluation metrics. Section 10 describes obstacles and open research problems. Section 11 discusses the practical applications and use cases. Section 12 discusses future research directions, while Section 13 closes the survey.

Related Work

Social networking sites have revolutionized the communication of people . Now , people may communicate with the world and share information . But this openness has also led to the production of hateful content that can hurt individuals and groups. In order to understand such material one requires precise meanings and differences between its various types. In this section, we cover the main principles relating to hostile posts and survey the forms of hostile content typically encountered on social media.

Key Definitions

- Hostile Post [12]: Content posted through social media insults, belittles, or threatens individuals or groups. Hostility can be in the form of text, visuals, or videos. This includes hate speech, cyberbullying, and insulting comments.
- Hate Speech [13] : A hostile post towards a person's innate qualities such as race, religion, or gender. It includes language that is racist or inflammatory that can lead to prejudice or violence . It is generally legally restricted .
- Cyberbullying [14]: Repeated harassment or intimidation of people through digital communication means. Cyber bullying is distinguished by its ongoing nature and its common victimization of vulnerable users. Cyber bullying can be expressed through threats, the dissemination of rumors or the sharing of private information.
- Offensive Language [15]: The use of profane or inappropriate terms that may not be directed at a person or group but contribute to a hostile climate online. Such language is often moderated to conform to community standards.
- Trolling[16]: Deliberate attempts to create disruption or argument by posting inflammatory or otherwise off-topic messages. Trolling may be meant as humor, but may also escalate into more serious activities such as hate speech.
- Misinformation and Disinformation [17]: Misinformation is the spread of misleading information without the intent to mislead, whereas disinformation is the spread of false information with the intent to mislead. Neither are antagonistic by nature but both can create enmity by creating mistrust and division.

Common Forms of Hostile Content

Hostile content on social media comes in many forms based on intent and context. Salminen et al. [18] identified Various types of hostile content:

- Direct Threats: Specific type of threat that causes bodily harm or violence with legal outcomes.
- Harassment and Personal Attacks [19]: Repeated personal attacks on individuals using abusive language, name-calling, or derogatory comments can lead to mental depression.
- Incitement to Violence or Hatred: Threats that produce discrimination or violence against groups based on identity criteria, with a major one causing significant societal impact.
- Defamation: False information or distorted posts aimed at damaging persons› reputations.

- Sexual Content [20]: Sexual or offensive content spread to offend particularly damaging when it has particularly harmful effects when directed at young users .

Distinctions Among Related Terms

It is helpful to differentiate between terms that are closely related terms.

- Hate Speech vs. Offensive Language [21]: Hate speech that promotes hatred, discrimination, or violence against individuals or groups based on protected characteristics such as race, religion, ethnicity, or gender. Offensive language, can be rude or vulgar without any discriminatory intent activity.
- Hate Speech versus Cyberbullying [22]: Hate speech is targeting groups of people based on race, gender or religion. while cyberbullying is against individuals and involves personal attacks over time.
- Trolling vs. Hate Speech [23]: Trolling means online destroying a piece of others by posting messages that are provoking or disruptive where as hate speech targets violence.
- Misinformation vs. unfriendly content [12]: Misinformation is incorrect information shared intentionally. whereas unfriendly content is intentionally used to criticise and upset. [8, 24, 25].

Methodological Framework

The present survey has been developed using a systematic and well-structured methodology aimed at organizing, interpreting, and evaluating existing approaches to hostile post detection on social media. The framework is based on three foundational pillars – model categorization, dataset exploration and analytical evaluation.

- Model Categorization: In order to achieve meaningful classification, hostile post detection methods are grouped into four broad classes: traditional machine learning, deep learning, hybrid systems and advanced methods. This taxonomy is based on underlying algorithmic structure and learning paradigm, helping to achieve clearer comparison of performance, scalability and adaptability.
- Dataset Exploration: A survey of popular public datasets is presented based on language variety, type of social media content (e.g., tweets, comments, video captions), topic domain and labelling approach. From this we identified a research gap that is bias during the annotation process, resulting in improper training of the model and lack of datasets in regional languages.
- Analytical Evaluation: Sections 3 to 9 provide a detailed survey of the literature. Each technique is evaluated for practical application, advantages and disadvantages, and evaluation techniques (e.g., accuracy, F1, ROC-AUC). This survey also contains challenges such as ethical implications, cultural sensitivity, and model transparency, crucial for building trustable systems. The future scope and scope of work also discuss advancements in existing techniques. It also highlights practical limitations that researchers face when tackling real-world hostile content across various platforms.

DATASETS FOR HOSTILE POST DETECTION

Overview of Existing Datasets Used for Hostile Post Detection

Table 1 summarizes the different datasets used for hostile post detection, indicating the language, type of

hostile content, size of the dataset, and a short description of each dataset. It points to resources for the study of hate speech, offensive language, and related phenomena across different languages and platforms.

Characteristics of datasets

In the analysis of datasets for hostile post detection, it is important to understand the following critical features for the applicability of a dataset for various research tasks. Here are the main features to keep in mind:

Properties of Datasets› Language:

- Language(s) in which the dataset is written is an important factor especially in multilingual and multicultural environments. Datasets can be monolingual (e.g., only English) or multilingual (e.g., English, Hindi, and Tamil). Few datasets are in the form of code-mixed language that contains mixed languages such as Hindi-English or English-Gujarati. Some (e.g., Hindi-English).
- Importance: The diversity of language creates problems in model generalisation due to different linguistic subtleties, slang, dialects and cultural contexts.
- Size: The size of the dataset is the number of records (e.g., tweets, comments, posts). The size of the dataset can be from thousands to millions of samples depending on the system scope and availability of data.

Importance: Large data sets are best for training machine learning models, but they can also present noise. A smaller dataset is properly annotated and well-maintained, but that unable to handle model generalisation.

- Annotation Process: The annotation is the process of labelling our dataset with properly defined categories. Annotations can be performed by subject experts, professionals or a combination of both. The annotation scheme can be binary classification or coarse grained classification (i.e. hate speech vs non-hate speech), multi-class classification or multi-label classification or fine grained classification (i.e. hate speech , importance The annotation process is heavily dependent on the quality and reliability of the dataset. More reliable datasets have explicit standards, more than one annotator to create redundancy, and high inter-annotator agreement. Some datasets even contain explanations or justifications for the corresponding label, which improves interpretability .

- Content Type: What is the type of content in the dataset? e.g., what platform it is from, e.g., Twitter, Facebook, or YouTube; what the content actually is, e.g., text, images, or video; and what it is about, e.g., news comments, social media discussions, or user reviews.

Importance: Social media allows spreading content in the form of text, images, video, GIFs, and emojis. This input dataset was collected and analysed based on text or multimodal data, as per the social media input dataset. Based on that, select a model and analyse data.

CHALLENGES ASSOCIATED WITH DATA COLLECTION AND ANNOTATION

Some of the challenges in data collection and annotation for hostile post detection include bias, cultural context, and data scarcity. Limited data is available for low-resource languages, so the model does not learn properly on few data. The annotation process is also completely depends on annotation knowledge and proper understanding of that domain. These problems are important to tackle to build accurate and fair hostile content detection systems [4].

Trends in Model and Dataset Usage

In addition to the dataset overview and modeling approaches discussed in prior sections, this subsection highlights two significant trends observed in recent studies on hostile post detection in social media. As shown in Figure 1, the utilization of transformer-based models (e.g., BERT and its variants) in the field is on the rise. The number of such papers has been increasing steadily from 2017 to 2024, which reflects the growing preference for context-aware deep learning models to handle unstructured social media data. Figure 2 Frequency of datasets used for detection of hostile content, grouped by language. The most common datasets are in English, followed by Hindi, and there are limited resources in regional languages like Tamil and Gujarati. The chart also includes multi-lingual datasets that are increasingly used for cross-lingual and code-mixed content analysis.

This section gives a comprehensive summary of publicly available datasets on hostile post detection. It provides a comparative summary of the main features, such as language, type of hostile content, and annotation procedure. It also gives issues related to quality of data, linguistic differences, and ethical problems. The section also explores trends in model usage and dataset preferences over time help to understand the latest model use trend in multilingual and transformer-based models.

FEATURE EXTRACTION TECHNIQUES

Feature extraction is critical in text processing for hostile post detection, converting raw text into numerical representations suitable for models. A summary and comparison of conventional and state-of-the-art feature extraction techniques are as follows.

- Introduction to Classical Feature Extraction Techniques

1. Bag of Words (BoW) [4]

– Description: BoW is an easy and most useful techniques for text feature extraction, where data is represented as a vector of word occurrences. It no focus on grammar and word order but only find the number of words.

– Pros: Methos is very easy to implement and suitable for simple text classification.

Drawbacks: It fail to preserve word order and syntax information, producing high-dimensional, sparse vectors.

2. Term Frequency–Inverse Document Frequency (TF-IDF) [4]: TF-IDF improves the result of the Bag of Words model (BoW) by adding weight to important words of a document.

– Description: TF-IDF enhances BoW by weighting the words of a document according to their importance to a collection. It is a combination of term frequency (TF) and inverse document frequency (IDF) .

– Pros: It adds importance to words in documents.

– Disadvantages: It unable to add order of words and context of larger datasets.

- Advanced Feature Extraction Techniques

1. Word Embeddings [42]

– Description: Word embeddings used to translate words into dense numerical vectors using feature extraction techniques like Word2Vec and GloVe.

– Advantages: It fetches semantic meaning and reduces dimensionality over BoW and TF-IDF.

– Disadvantages: Static embeddings do not capture word senses and have problems with unknown words.

2. BERT-Based Embeddings [43]:

- Description: Bidirectional Encoder Representations from Transformers (BERT) understand detailed text in a bidirectional way and give perfect contextual embeddings, capturing all hidden information.

– Pros: Capable of identifying surrounding context and hidden feature extraction and also capable of model fine-tuning based on real-world application.

– Cons: It requires high computational power and proper fine-tuning to achieve the best model performance.

This is a detailed comparative study of various feature extraction techniques specifically for detecting hostile posts.

1. Accuracy and Performance:

– Traditional Methods: These methods are easy to implement and computationally cheap but unable to understand the context of hostile posts in detail.

– Advanced Methods: Transformer-based models are capturing semantic meaning and contextual cues, which gives the best performance.

2. Computation efficiency:

– Traditional Methods: The model performs well on small datasets, and also it requires less computation power.

Table 1. Comparison of datasets for hostile post detection across languages and content types

Dataset Name	Language	Type of Hostile Content	Size	Description
Hate Speech and Offensive Language Dataset (HSOL) [26]	English	Hate Speech, Offensive Language	24,783 tweets	On the other hand, annotations of hate speech and abusive language on tweets are often used for constructing and evaluating models.
Stormfront Data [27]	English	Hate Speech, Extremism	10,000+ posts	collected from the Stormfront forum, focusing on white supremacist content in hate speech.
Facebook Hate Speech Dataset [28]	Italian	Hate Speech	5,000 comments	Facebook comments labeled for hate speech, examining its frequency in the Italian language.
HASOC Dataset [29]	English, Hindi, German	Hate Speech, Offensive Language	20,000+ tweets	Multilingual dataset for hate speech and offensive language identification
Cyberbullying Dataset [30]	English	Cyberbullying	47,000+ instances	Social media data with annotations for different types of cyberbullying and aggression post.
SentiStrength Twitter Dataset [31]	English	Offensive Language, Insults	6,000 tweets	Designed for identifying offensive language and insults for sentiment analysis.
GermEval 2018 Dataset [32]	German	Offensive Language	9,000+ tweets	Created for offensive language detection, with annotations for binary and multi-label classification.
TRAC-2 Dataset [33]	English, Hindi, Bengali	Aggression, Hostility	15,000+ posts	Provides comments annotated for different levels of aggression and hostility for the TRAC shared task.
Hate Speech and Offensive Content (HateXplain) [34]	English	Hate Speech, Offensive Language, Target ID	20,000+ tweets	collected dataset with hate speech classes and accompanying explanations for better model interpretability.
Danish Abusive Language Dataset (DALC) [35]	Danish	Abusive Language, Hate Speech	3,000 comments	Danish social media comments annotated for abusive language and linguistic subtleties.
Dravidian Code-Mix Dataset [36]	Tamil, Malayalam	Offensive Language, Hate Speech	12,000+ comments	Code-mixed comments from YouTube annotated for offensive language, part of the HASOC-Dravidian work.
Hate Speech Detection in Hindi-English Code-Mixed Social Media Text [37]	Hindi-English (Code-Mixed)	Hate Speech, Offensive Language	15,000+ tweets	Dataset of code-mixed tweets labelled for detecting hate speech and offensive language.
Offensive Language Identification Dataset in Indian Languages (OffensEval-2020) [38]	Hindi, Bengali, Tamil	Offensive Language, Hate Speech	30,000+ posts	Posts in multiple Indian languages labelled for offensive language as part of the OffensEval work.
Gujarati Offensive Language Dataset (GOLD) [39]	Gujarati	Offensive Language, Hate Speech	5,000+ tweets	labelled Gujarati posts for offensive language and promoting NLP research for low-resource languages.
Kannada Offensive Speech Dataset (KOSD) [40]	Kannada	Offensive Language, Hate Speech	2,500+ comments	A large dataset for detecting offensive speech in Kannada and labelled with related content.
Gujarati Hostile Post Detection Dataset [4]	Gujarati	Hostile Post	10,000 posts	Created the Gujarati language hostile post detection (GUHPD) dataset with binary and multi-class labels for hostility detection
Indic Offensive Content Dataset [41]	Hindi, Bengali, Tamil, Kannada, Malayalam	Offensive Language, Hate Speech	50,000+ posts	A dataset in multiple Indian languages annotated for offensive posts to do multilingual NLP research.

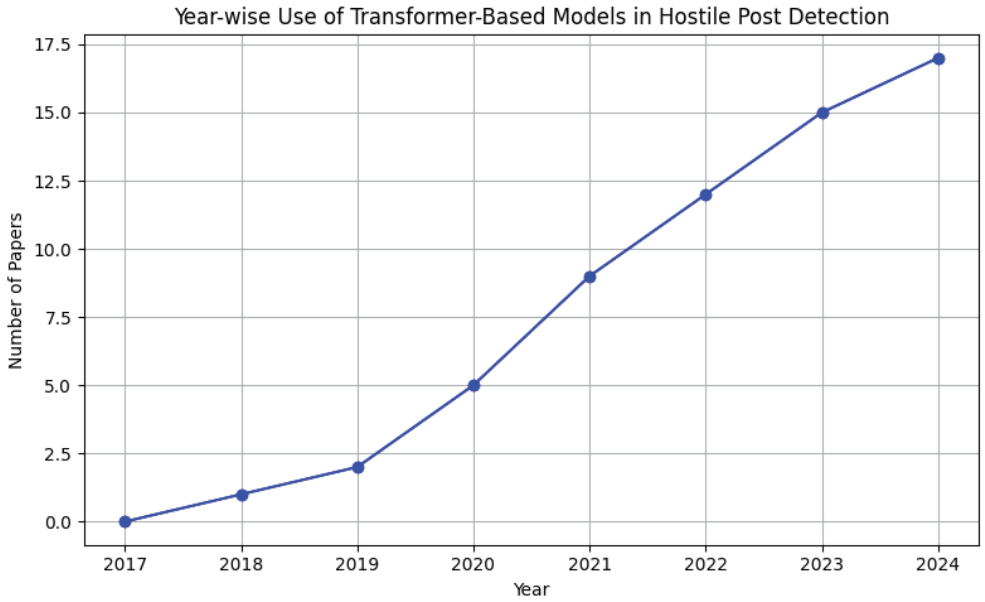

Figure 1. Year-wise use of transformer-based models in hostile post detection (2017–2024).

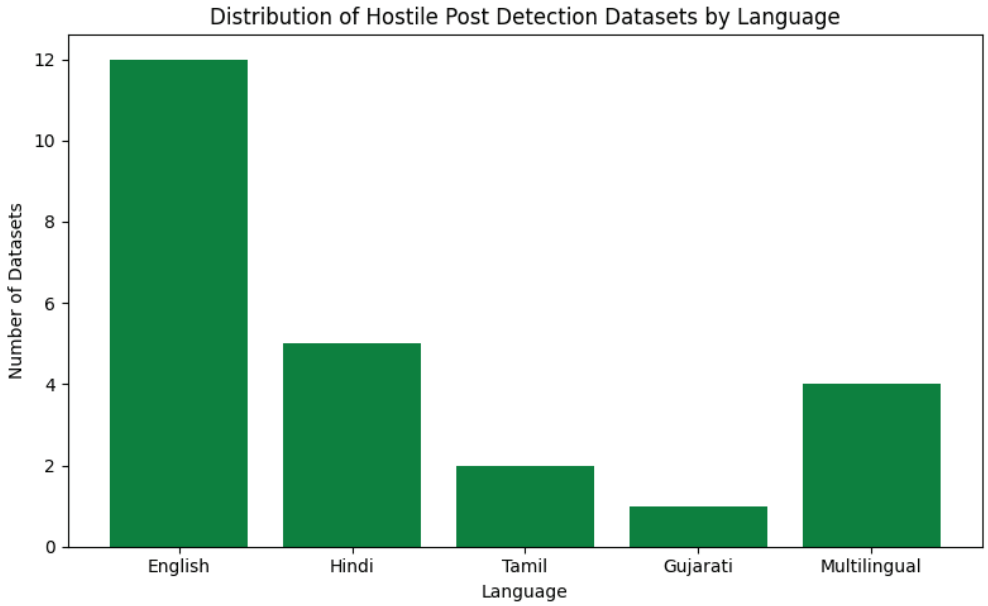

Figure 2. Distribution of hostile post detection datasets by language.

– Advanced Methods: The Transformer-based BERT model and advanced model are powerful for capturing the context of a sentence, but they also require high computation power.

3. Grasping Context and Semantics:

– Traditional Methods: Model unable to handle hostile language detection if it is indirect intent.

-Advanced Methods: They capture semantic meaning insight, which is necessary to accurately detect context-sensitive hostile posts from social media.

4. Flexibility and Adaptability:

– Traditional Methods Less flexible to newly added data, they need retraining for newly added words.

– Advanced Techniques: Advance contextual models capable of providing flexibility, such as the ability to fine-tune for various other areas or languages.

This section summarises traditional and advanced feature extraction methods used in hostile post detection. Techniques like BoW and TF-IDF are easy and

computationally efficient but lack the ability to capture detailed semantic meanings. Embedding-based models, especially the Transformer model BERT, are best for capturing contextual meaning and are well-suited for large-scale datasets and multilingual social media content monitoring.

Machine Learning Approaches

Machine learning-based methods are fundamental techniques for detecting hostile posts. Support Vector Machines (SVMs) perform best in high-dimensional spaces and classify text by capturing the optimal hyperplane that separates various classes. Naive Bayes classifiers use the concept of Bayes' theorem and assume that features are independent. They provide the best results for text classification, particularly for detecting spam or offensive language, even with large datasets. Random Forest is an ensemble method that constructs multiple decision trees and combines their predictions. It enhances classification accuracy and mitigates overfitting through majority voting or averaging. Learning methods vary depending on the availability of labeled data. Supervised learning required labeled datasets, enabling models to learn from input-output pairs. It is most effective when there is a sufficient amount of labeled data. However, unsupervised learning techniques, able to capture patterns without labeled data, may be able to identify overview of structures but may not be as precise in capturing hostile post. Semisupervised learning combines labeled and unlabeled data, taking advantage of the abundance

Table 2. Comparison of machine learning approaches for hostile post detection

Machine Learning Algorithm	Description	Advantages	Disadvantages
Support Vector Machine (SVM) [44]	SVM can find the optimal hyperplane to divide various classes in the feature space, which is suited for high dimensional data.	Good in high dimensional context, performs well for binary classification, less prone to overfitting.	Parameters require careful adjustment ; sluggish on large datasets ; may struggle with noisy data .
Naive Bayes [45]	This algorithm relies on Bayes› theorem, assuming that features are independent, and is particularly effective for classifying text.	Easy to implement and computationally light; works well with extensive datasets; suitable for text with many features.	unable to handle complex language pattern.
Random Forest [46]	It is Ensemble based approach. Merge multiple decision trees and combine their prediction accuracy.	It handles all types of data and simple and efficient model.	Required proper hyperparameter fine-tuning with large resource utilisation.
Logistic Regression [47]	Use a logistic function to predict probability and best suitable for binary classification on linear data.	Simple to easy to implement and suitable for small dataset.	unable to learn complex data and large dataset.
K-Nearest Neighbors (KNN) [48]	The algorithm classifies a point based on the majority class of its k nearest neighbors.	Easy to implement and intuitive; flexible; no training phase required.	May be computationally intensive on large datasets; sensitive to choice of k and distance metric; may struggle with high dimensional data.
Decision Trees [49]	A model provide decisions by dividing data based on values of feature extraction.	Works bets for numerical and categorial data and doesn›t require feature scaling	It fit to training data but does not perform well on unseen data.
Gradient Boosting Machines (GBM) [50]	An ensemble method that builds models sequentially that focus on correcting mistakes of previous model.	Useful for complex data structures and capable of finding hidden patterns.	Efficiently provide carefully. Change results on noisy data and large resources required for computation.
AdaBoost [51]	A boosting algorithm improve the performance of weak classifiers by focusing on the incorrect classification made by previous models.	Overfitting reduce and improve performance n complex problems.	Performance decrease on noisy data; requires careful parameter fine tuning; can overfit without proper control.
Word Embeddings (Word2Vec, GloVe) [52]	These methods convert words into dense vectors, capturing their meanings and relationships.	Captures the semantic relationships between words; reduces dimensionality; offers improvements over traditional methods like Bag of Words.	Static embeddings may not effectively address polysemy; they require pre-trained models or substantial computational resources.
BERT-Based Embeddings [52]	Utilizes transformer models to give contextualized word embeddings, capturing detailed semantic information based on context.	Highly effective for understanding complex language ruffness; context-aware embeddings enhance accuracy.	Computationally intensive; requires high resources for model training and fine-tuning; implementation can be complex.

of unlabeled data when labeled data is limited and often leads to better generalization than supervised learning alone. Ensemble techniques, such as bagging and boosting, improve performance by combining predictions from several models. Bagging uses the average of the predictions of models trained on different data subsets to reduce variance, while boosting minimizes residual errors by sequential training, leading to a robust model. Each approach has its advantages and disadvantages. Traditional methods like SVM and Naive Bayes are good at handling smaller datasets but may not be suitable for complex, high dimensional data. Ensemble methods are powerful but computationally expensive. Supervised learning methods are accurate but heavily reliant on the availability of labeled data. Unsupervised methods are flexible but may lack accuracy. Semi-supervised learning methods are a good compromise between the need for labeled data and the ability to utilize unlabeled data but necessitate good data management.

In this section, traditional machine learning techniques for hostile post detection were presented. Their learning paradigms, strengths and limitations were explained. Algorithms like SVM and Naive Bayes are simple and efficient but ensemble methods give better accuracy at the cost of high computation. The comparative table summarizes their application cases that help the researchers to choose appropriate models based on the type, size, and complexity of the data (Table 2).

Deep Learning and Neural Network-Based Approaches

Significant progress has been made in the detection of hostile posts using deep learning and neural network-based techniques that recognize complicated patterns in textual data by employing sophisticated architectures . The most basic form of neural network is the Feedforward Neural Network (FNN). Here the data flows in one direction only, from the input layer, through the fully connected layers and to the output layer. These networks are easy to implement and perform well for small to medium sized data sets capturing non-linear relationship in an efficient way. However they have difficulties with sequential or contextual information, and are therefore less suitable for tasks that require understanding temporal dependencies. Convolutional Neural Networks (CNNs) have been designed for image analysis, but have shown great potential for text classification as well. Convolutional layers enable CNNs to learn local patterns and features in the text, thus helping to handle different input lengths and decreasing dimensionality. They are good at capturing local features, but they often fail to capture long-range

Table 3. Comparison of deep learning approaches for hostile post detection

Deep Learning Approach	Description	Advantages	Disadvantages
Pre-trained Transformers (e.g., BERT, GPT) [53]	These models are transformer-based model that trained on large datasets and fine-tuned for applications.	Achieves excellent performance; provides rich contextual understanding; reduces extensive model training needs.	Requires more computational power and complexity for fine-tuning hyperparameter adjustment.
Feedforward Neural Networks (FNN) [54]	Basic neural networks with fully connected layers, where data flows in one direction.	It is easy to implement, suitable for medium datasets and also capable of understanding non-linear patterns.	Sequential data and large data is not handling properly.
Convolutional Neural Networks (CNN) [23]	Effectively identify local patterns in both text and image data.	It is efficient in identifying patterns on varying lengths of data.	Layer tuning is required properly and not handle long-range dependencies.
Recurrent Neural Networks (RNN) [12]	It›s designed to handle sequential data; also, it properly remembers information from earlier input.	It captures information from previous inputs and also remembers information from earlier input.	Training is slow because gradient process hinder effective learning
Long Short-Term Memory Networks (LSTM) [55]	An improved RNN by adding memory cell helps to understand long term dependencies	Effectively captures long-range dependencies; alleviates gradient issues.	Resource-intensive; complex; may struggle with very long sequences.
Gated Recurrent Units (GRU) [15]	A lightweight LSTM model captures long-term sequences using fewer resources.	Less resource use for solve the problem.	It not perform well for complex long sequence.
Transformer Networks [23]	Use a self-attention mechanism to assess the importance of different words within sequences.	Effectively captures context and is also scalable for long dependencies handling.	High computational resources are required, complex for implementatio and needs extensive fine-tuning.

dependencies, which may affect performance in contextually rich tasks. Then came Recurrent Neural Networks (RNNs) that have directed cycles that allow them to remember and interpret temporal dependencies, allowing them to process sequential data. This makes RNNs well suited to tasks where the order of the input matters. However, they suffer from problems such as vanishing and exploding gradients which may affect the effectiveness of training, especially for long sequences . Long Short-Term Memory Networks (LSTMs) are an extension of RNNs with additional memory cells and gating mechanisms that allow them to better deal with long term dependencies. Long short-term memory networks can solve the vanishing gradient problem. They are good for tasks with long sequences and complex contextual dependencies. But they are computationally intensive and complex to implement, which can be a hurdle in resource-constrained environments. GRUs (Gated Recurrent Units) are a simpler variation of LSTMs by cutting down the number of gates to two. This lowers the computational needs while still being able to effectively capture sequential dependencies. GRUs are more efficient than LSTMs, but they may not perform as well on long sequences and complex tasks. Transformer networks represent a significant advancement in neural network architectures, incorporating self-attention mechanisms that assess the importance of different words in a sentence to each other, enabling a more comprehensive approach to contextual relationships. They are adept at long-range dependencies; they are scalable and hence applicable to complex applications. But they are resource-hungry and require much fine-tuning. Such drawbacks could be a barrier for practical use. Pre-trained transformer models, such as BERT and GPT, trained on large data sets, have been successfully used for a variety of natural language processing tasks, including hostile post-detection. They are rich in context and cut down the need for large, task-specific training significantly. But they are expensive to calculate and difficult to implement. System need a high level of computation resources and perfect fine-tuning to perform well on particular datasets. Generally, deep learning and neural network-based approaches are powerful for hostile content detection, and each has its pros and cons. They include FNNs and complex pre-trained transformers with different functionalities for text analysis and successfully identify hostile posts. Table 3 summarizes the different deep learning techniques for hostile post detection along with their main features, advantages, and disadvantages. This comparison will help to understand the most appropriate ways of handling the various kinds of tasks and data.

This section gives a detailed summary of deep learning techniques for hostile post detection, ranging from basic neural network models to advanced pre-trained transformer models. While FNNs, CNNs, and RNNs provide basic architectures, models like LSTM, GRU, and Transformers outperform them in contextual and semantic understanding. The comparative survey table highlights each model's trade-offs in terms of accuracy, efficiency and complexity, guiding the selection of appropriate techniques based on application needs.

HYBRID AND ADVANCED TECHNIQUES

Hybrid and advanced algorithms for hostile post identification integrate multiple strategies to boost performance and address the limits of individual methods. Some of the hybrid methods are ensemble methods, which is a family of hybrid methods that combines predictions from numerous models to boost overall accuracy. Bagging (bootstrap aggregating) reduces both the variance and the overfitting (especially on noisy data sets) by building numerous models on random subsets of the training data and averaging their predictions. Another is boosting . This focuses on hard cases and incrementally trains models to fix the flaws of the previous models . It is used by algorithms like Gradient Boosting Machines (GBM) and AdaBoost to create powerful classifiers for complex data. Stacking is a way to improve the predictive performance of a meta-model by combining predictions of a large number of models, thus taking advantage of the benefits of different base models. The multiplicity of models is good for voting, and makes it more resilient by combining predictions with majority voting or average. Advanced Techniques are creative ways to improve the model performance. Transfer learning uses the knowledge from large pre-training models such as BERT or GPT to fine-tune for specific tasks and to get good performance with smaller datasets. Meta-learning aims to build models that can adapt quickly to new tasks with few data. This means that models for hostile post detection can be changed for multiple languages or platforms. cAttention Mechanisms, especially those in transformers, help the model to focus better on the essential parts of the input data and thereby improve the contextual knowledge. Some of the methods that have been developed to enable models to attend to different parts of a text include Self-Attention and Multi-Head Attention . This allows for subtle inferences about hostile material. Table 4 Summary of some advanced strategies in hostile post detection. Method Description Advantages Drawbacks Bagging Classifier ensemble method . It can help improve the overall accuracy . It is computationally expensive . Boosting Classifier ensemble method . It can help improve the overall accuracy . It is computationally expensive . Transfer learning Transfer knowledge acquired in one domain to another domain . It can help improve the overall accuracy . This section discussed hybrid and advanced strategies used to improve the performance of the hostile post detection. Ensemble techniques such as Bagging, Boosting and Stacking integrate several models to boost the robustness and accuracy. Advanced techniques such as transfer learning and meta-learning enhance adaptability, while attention processes permit deeper contextual analysis. The trade-offs of the comparative table are shown for a suitable decision according to the computational power and tasks complexity.

Table 4. Overview of advanced techniques for hostile post detection

Technique	Description	Advantages	Drawbacks
Bagging (Bootstrap Aggregating) [56]	Constructs several models using random subsets of data and combines their predictions through averaging.	Reduces variance; improves robustness; prevents overfitting.	It required higher computation power and not perform well when base model perform poor.
Boosting (e.g., AdaBoost, GBM) [57]	It improves model performance in sequence and corrects mistakes of earlier model.	Capable of handling complex data and reducing bias.	Computational power required and cannot handle noisy data.
Stacking [58]	Combines predictions from different models via a meta-model to produce a final output.	Uses strengths of different base models; improves prediction performance.	Difficult to implement; requires careful selection of base models and meta-models.
Voting [59]	Combines predictions of different models by majority voting or averaging.	Easy to implement; leverages model diversity; improves robustness.	Does not necessarily improve performance if base models are similar; it may require careful weighting.
Transfer Learning [60]	A pre-training model was used and fine-tuned for a specific task.	It performs excellently on small datasets because we are using a pre-trained model on large dataset.	High computation resources are essential for model building.
Meta-Learning [61]	Create model that can learn fast on new and minimal data	fast learning to new tasks; useful in changing environments and variable data.	Requires advanced techniques; may not perform well with limited dataset.
Attention Mechanisms [62]	Apply self-attention mechanisms to highlight the importance of words.	Improves contextual understandingof data; effective for complex data.	Computationally intensive; complex to implement and tune.

Comparative Summary of ML, DL, And Hybrid Approaches

In this section, we provide a comparative analysis of hybrid, deep learning (DL) and machine learning (ML) approaches for the identification of aggressive communications on social media, based on previous talks on specific paradigms. This section attempts to integrate the rich examination of the concepts and distinctive contributions of each technique, focusing on the practical trade-offs across a spectrum of critical dimensions. Table 5 shows the characteristics of different methods regarding their effectiveness, compatibility with datasets, scalability, and support for multilingual data. This tabular overview helps researchers select the most appropriate technique for their use case by assessing technical limitations and application requirements. It shows that traditional machine learning algorithms are scalable by extending hardware capacity and

Table 5. Comparative study of ML, DL, and hybrid techniques for hostile post detection

Technique Type	Performance	Dataset Types	Scalability	Multilingual Support
Machine Learning (ML)	The result provides the best for structure and small datasets. Not suitable for noisy and large data.	Manually feature extraction required and structure label data is necessary for model.	Scalability is possible by extending hardware resourcing.	It is not suitable for multilingual datasets, and its performance decreases if we do not provide language-specific features.
Deep Learning (DL)	Easily capture context and provide the best accuracy for the detection of hidden features and semantic relationships between data.	The model can detect unstructured , noisy data. It does not require manual feature extraction.	Need high computational resources and efficient results provided by use of GPU and TPU computation power,	Supports multilingual input through models like mBERT, XLM-R. Capable of zero-shot and cross-lingual inference.
Hybrid Techniques	ML and DL combination of models provides best and excellent result on various datasets for different class distributions.	It is capable of handling both structure and unstructuure data.	Model scalability depends on model design, combining multiple models› complexity.	When we provide pretrained model and efficient preprocessing, then it provide efficient multilingual hostile post detection support

simple, but they are not sufficient for unstructured, multilingual or complex data. However, deep learning models have a high computing cost, but they have greater efficacy. Hybrid algorithms are designed to find a middle ground among these trade-offs and often perform better at detecting malicious information in the actual world.

EVALUATION METRICS

The hostile post detection model is evaluated using various evaluation parameters. The evaluation score provided by the more accurate metrics is:

- Accuracy [63]: It gives the percentage of correctly classified posts (both hostile and non-hostile) out of the total number of posts.

$$Accuracy = (True\ Positives\ (TP) + True\ Negatives\ (TN)) / Total\ Instances \quad (1)$$

Accuracy is easy to calculate and understand, but it does not provide the best result for an imbalanced dataset.

- Precision [64]: This metric measures the correctly identified positive posts among all posts predicted as positive by the model.

$$Precision = \text{True Positives (TP)} / (\text{True Positives (TP)} + \text{False Positives (FP)}) \quad (2)$$

- Recall [65]:

‹Recall› represents the actual positive posts that are truly predicted as positive by the model.

$$\text{Recall} = \text{True Positives (TP)} / (\text{True Positives (TP)} + \text{False Negatives (FN)}) \quad (3)$$

- F1 Score [65]: It combines the measures of precision and recall together and gives a balanced measurement of the model.

$$\text{F1 Score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

- AUC-ROC (Area Under the Receiver Operating Characteristic Curve) [64]: This metric checksposts how the model differentiates hostile and non-hostile posts based on the threshold value.

$$AUC\text{-}ROC = \int_{-\infty}^{\infty} ROC\ Curve \quad (5)$$

A higher AUC-ROC value indicates superior performance in accurately classifying instances.

Challenges in Evaluating Hostile Post Detection Models

Evaluation of hostile post detection algorithms is challenging for the following reasons

– Imbalance in Datasets [66]: There are usually more hostile postings than non-hostile postings. This discrepancy can skew metrics like accuracy, making models appear better than they really are. In these cases, measures like precision, recall and F1-score gives a better view, focusing on performance with regard to minority class.

– Contextual and Cultural Variability [67]: The concept of hostile content can differ widely across cultures and circumstances. This could make a dataset that works well not function well with a model trained on another dataset since contextual details may change. Diversity is necessary for generalization of evaluation measures and demands context-sensitive judgements.

- Dynamic Nature of Language [68]: Language changes affect model performance, so retraining is required over a period of time for improved performance of model.
- Subjectivity in Annotation [69]: The classification of posts as hostile or non-aggressive is a subjective problem. It can cause errors in ground truth data. The subjectivity of the evaluation measures influences their reliability and requires careful annotation methods to ensure correctness and consistency. Such issues demonstrate the requirement of large and adaptable evaluation frameworks to fit for the intricacies of hostile post detection.

DISCUSSION ON THE NEED FOR NEW OR IMPROVED METRICS

Hostile post identification algorithms need new or improved measures to evaluate them.

- Context-Aware Metrics [70]: Metrics considering the contextual and cultural characteristics can be useful to evaluate models in diverse situations. More context-sensitive measures would be more appropriate for capturing the subtlety of what constitutes detrimental content in diverse circumstances.
- Dynamic Metrics [70]: Flexible metrics that evolve with language and context through time may help keep models relevant. Evaluation methods can learn from new data and adjust thresholds based on continual learning.
- Granular Metrics [71]: Separate metrics used to identify the detail of the category (e.g., hate speech vs cyberbullying) of data accurately. That also helps fine-tune the model based on their category result.
- Multi-Aspect Evaluation [72]: A combination of traditional and advanced evaluation metrics is used to better evaluate the model's performance. Traditional evaluation metrics such as accuracy, precision, recall, F1-score, and AUCROC are basic, but they are not enough for complete hostility detection.

CHALLENGES AND OPEN ISSUES

Data-Related Challenges

Hostile post detection has numerous key data challenges. One key difficulty is the imbalance of dataset as there are far less hostile than non-aggressive posts. This results in a biased performance measurement. This mismatch can bias the accuracy measures and predispose the

models to the majority class . Solutions include over-sampling the hostile posts, under-sampling the benign postings or employing synthetic data methods such as SMOTE to balance the data set . Multilingual detection is a more hard problem for the models because they must perform well across languages that have diverse linguistic and cultural characteristics. This comprises strategies such as building language-specific models or using multi-lingual embeddings and transfer learning for better adaption. The annotation problems are also attributable to the subjectivity of the decision of whether a post is hostile or not, which may lead to discrepancies. To tackle this problem it is important to set tight annotation standards and utilize numerous annotators to make sure that there is consensus and consistency [73–75].

Model-Related Challenges

We discuss the problems of overfitting, generalization and interpretability for the hostile post detection model. This is called "overfitting". The models overfit the training data and learn the noise that does not generalize well to new data . This leads to poor performance in the real world. Cross validation and regularization can help with generalization . Another challenge is how to generalize the models well to various datasets. Models trained on a dataset may not generalize to another dataset owing to language or cultural variations. Training on several datasets helps generalization . Also , domain adaptation or transfer learning can aid . Finally, interpretability is important especially when complicated models are used as "black-box". To increase transparency, it is necessary to employ approaches that can comprehend the model's results such as feature importance analysis and explainable AI techniques.

Ethical and Social Challenges

Ethical and social issues: bias, privacy and regulatory impediments Bias Bias is a concern as models may have assumptions based on the data they are trained on and may provide biased results in terms of race, gender, socioeconomic class etc. Fair-aware algorithms and diverse and representative data sets may help to alleviate them.We do care about privacy and specifically personal data and sensitive businesses. Privacy must be anonymized and controlled by law. law:Content filtering is a difficult wicket when you see legislation in other domains. What are the data protection rules we have to follow to be ethical ? Holistically addressing these ethical and sociological factors requires technical solutions focusing on fairness, privacy and legal compliance to enable responsible hostile post detection [76,77].

Applications and Use Cases

Hostile post detection technology has several essential uses. It is used to monitor user-generated content on social media sites, and to detect and flag harmful posts to create a safe online environment. These systems automate moderation tasks, and lessen the load on human moderators. These tools fight crime by analyzing interactions happening online. In the policy realm, these systems provide a window into online behaviors that shape the development of rules and initiatives for online safety. Another important area is cross-lingual detection, which enhances the ability to detect hostile content around the globe. This feature is vital for global companies that deal with online hostility in different cultural settings [78, 79].

Future Directions and Research Opportunities

Hostile post detection is a developing field and there are many possibilities for further investigation. Major trends include multi-modal analysis, which combines text, graphics and video to improve content identification by offering more context. The accuracy and contextual comprehension of transformer-based models and pre-trained language models are improving. A critical area to work on for the multilingual data challenges is the development of cross-lingual models that work well for detecting hostile information in multiple languages . Another interesting area is to enhance multi-modal analysis to incorporate textual and visual data. This can assist models to better understand context and detect subtle hostile content more accurately. There is also increasing need for additional varied datasets that are representative across multiple languages and cultures. What is needed are fresh models with fairness-aware algorithms that are able to adapt to changing linguistic patterns. Further, the development of ethical frameworks will ensure that hostile post detection tools are used responsibly. Future studies should seek to take use of these tendencies to improve the detection capabilities, while maintaining ethical norms in the field [79, 80].

CONCLUSION

The present survey has offered a thorough examination of techniques for detecting hostile content on social media, an area that continues to gain importance due to the proliferation of online abuse, misinformation, and harmful language. Through a review of traditional machine learning methods, advanced deep learning architectures, and hybrid approaches, we have highlighted the strengths and limitations of each category. In our analysis we find that while technical progress has significantly enhanced detection capabilities, there remain several challenges to overcome. These include the scarcity of data in low-resource languages, the skewed distributions in datasets, and the ethical considerations of automatic moderation. These challenges are both a technical and a societal imperative. Among the main findings of this research is a rising need for multilingual, multimodal approaches especially in culturally diverse digital environments . Equally important is the need for fair and transparent models that foster trust between end-users, policy-makers and platform managers. As online communication continues to evolve, so must the tools we employ to safeguard digital spaces. Future work should thus be directed towards the construction of inclusive datasets,

adaptive learning models and ethical frameworks that will be able to accommodate the intricacies of real-life situations. The combination of reinforcement learning, meta-learning and cross-lingual embeddings is a particularly promising direction for achieving this goal. In conclusion, this review aims at being a reference point for researchers and practitioners alike. By pointing out the existing gaps and emerging trends, we hope to stimulate more informed, ethical and technically sound research in hostile post detection.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Naslund JA, Bondre A, Torous J, Aschbrenner KA. Social media and mental health: Benefits, risks, and opportunities for research and practice. J Technol Behav Sci 2020;5:245–257.

[2] Jain AK, Sahoo SR, Kaubiyal J. Online social networks security and privacy: Comprehensive review and analysis. Complex Intell Syst 2021;7:2157–2177.

[3] Chakraborty A, Roy S. An ensemble model for hostile content detection in regional languages. J Comput Soc Syst 2024;12:33–50.

[4] Boda JR, Rana K. Investigating traditional machine learning approaches for hostile post detection. Int J Cyber Psychol Behav 2024;3:88–102.

[5] Chanda P, Verma R. Towards context-aware deep learning for detecting abusive language in code-mixed text. AI Soc 2024;39:115–129.

[6] Sharma P, Mehta A. Abusive language detection using BERT and BiLSTM in low-resource languages. Int J Artif Intell Res 2024;18:211–228.

[7] Nikitha R, Rao KP. Fake news and abusive content detection using hybrid semi-supervised frameworks. Procedia Comput Sci 2022;200:1431–1440.

[8] Mishra SU, Mishra LN, Mishra RK, Patnaik SR. Some applications of fractional calculus in technological development. J Fract Calc Appl 2019;10:228–235.

[9] Mohanta KK, Sharanappa DS, Dabke D, Mishra LN, Mishra VN. Data envelopment analysis on the context of spherical fuzzy inputs and outputs. Eur J Pure Appl Math 2022;15:1158–1179.

[10] Gaffar A, Joshi AB, Singh S, Mishra VN, Rosales HG, Zhou L, et al. A technique for securing multiple digital images based on 2D linear congruential generator, silver ratio, and Galois field. IEEE Access 2021;9:96125–96150.

[11] Sharma MK, Dhiman N, Mishra LN, Mishra VN, Sahani SK. Mediative fuzzy extension technique and its consistent measurement in the decision making of medical application. Math Probl Eng 2021;2021:1–14.

[12] Bhatnagar V, Kumar P, Bhattacharyya P. Investigating hostile post detection in Hindi. Neurocomputing 2022;474:60–81.

[13] Roy PK, Tripathy AK, Das TK, Gao XZ. A framework for hate speech detection using deep convolutional neural network. IEEE Access 2020;8:204951–204962.

[14] Iwendi C, Srivastava G, Khan S, Maddikunta PKR. Cyberbullying detection solutions based on deep learning architectures. Multimed Syst 2023;29:1839–1852.

[15] Hajibabaee P, Malekzadeh M, Ahmadi M, Heidari M, Esmaeilzadeh A, Abdolazimi R, et al. Offensive language detection on social media based on text classification. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE; 2022. p. 92–98.

[16] Uyheng J, Moffitt J, Carley KM. The language and targets of online trolling: A psycholinguistic approach for social cybersecurity. Inf Process Manag 2022;59:103012.

[17] Faustini PHA, Covoes TF. Fake news detection in multiple platforms and languages. Expert Syst Appl 2020;158:113503.

[18] Salminen J, Hopf M, Chowdhury SA, Jung SG, Almerekhi H, Jansen BJ. Developing an online hate classifier for multiple social media platforms. Hum Cent Comput Inf Sci 2020;10:1–34.

[19] Pendergrast TR, Jain S, Trueger NS, Gottlieb M, Woitowich NC, Arora VM. Prevalence of personal attacks and sexual harassment of physicians on social media. JAMA Intern Med 2021;181:550–552.

[20] Udupa S, Gerold OL. "Deal" of the day: Sex, porn, and political hate on social media. In: Social processes of online hate. London: Routledge; 2024. p. 120–143.

[21] Mandl T, Modha S, Kumar MA, Chakravarthi BR. Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation. 2020. p. 29–32.

[22] Sultan D, Toktarova A, Zhumadillayeva A, Aldeshov S, Mussiraliyeva S, Beissenova G, et al. Cyberbullying-related hate speech detection using shallow-to-deep learning. Comput Mater Contin 2023;74:2115–2131.

[23] Bhatnagar V, Kumar P, Moghili S, Bhattacharyya P. Divide and conquer: An ensemble approach for hostile post detection in Hindi. In: Combating online hostile posts in regional languages during emergency situation: First International Workshop, CONSTRAINT 2021, collocated with AAAI 2021, Virtual Event, February 8, 2021, revised selected papers. Springer; 2021. p. 244–255.

[24] Mohanta KK, Sharanappa DS, Dabke D, Mishra LN, Mishra VN. Data envelopment analysis on the context of spherical fuzzy inputs and outputs. Eur J Pure Appl Math 2022;15:1158–1179.

[25] Gaffar A, Joshi AB, Singh S, Mishra VN, Rosales HG, Zhou L, et al. A technique for securing multiple digital images based on 2D linear congruential generator, silver ratio, and Galois field. IEEE Access 2021;9:96125–96150.

[26] Hate Speech and Offensive Language Dataset. Hate speech and offensive language dataset. 2020 Available at: https://paperswithcode.com/dataset/hate-speech-and-offensive-language Accessed on Jul 4, 2024.

[27] Williams L. Stormfront hate speech dataset. 2018. Available at: https://example.com/stormfront-dataset Accessed on Jul 4, 2024.

[28] Rossetti S. Facebook hate speech dataset. 2018 Available at: https://example.com/facebook-hate-speech-dataset Accessed on Jul 4, 2024.

[29] Zampieri M, Lee J. Overview of the HASOC 2018 shared task on hate speech and offensive content identification. In: Proceedings of the 5th Workshop on Asian Language Resources and Evaluation. 2018. p. 1–7.

[30] Cyberbullying Dataset. Cyberbullying dataset [Internet]. 2021 [cited 2024 Jul 4]. Available from: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset

[31] Thelwall M, Harries P. SentiStrength: A simple tool for measuring the strength of sentiment in text. In: Proceedings of the International Conference on Weblogs and Social Media. 2014. p. 1–8.

[32] Wiegand M, Klakow D, K W. Overview of the GermEval 2018 shared task on offensive language. In: Proceedings of the GermEval 2018 Workshop. 2018. p. 1–10.

[33] Mall D, Ghosh S, Naik P. TRAC-2: Dataset and baseline for aggression detection in social media. In: Proceedings of the 12th International Conference on Language Resources and Evaluation. 2020. p. 105–112.

[34] Borkan D, Dixon L, Sorensen J. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency. 2019. p. 1–15.

[35] Nielsen A. Danish abusive language dataset. 2019. Available at: https://example.com/danish-abusive-language-dataset Accessed on Jul 4, 2024.

[36] Choudhury M, S M. Dravidian code-mix dataset. In: Proceedings of the 2021 Shared Task on Code-Mixed Text. 2021.

[37] Sharma S, Kumar A. Hate speech detection in Hindi-English code-mixed social media text. Int J Data Sci Anal 2021;12:341–356.

[38] Reddy M, D S. Offensive language identification dataset in Indian languages (OffensEval-2020). In: Proceedings of the 14th Conference on Natural Language Engineering. 2020. p. 45–56.

[39] Patel J, Singh R. Gujarati offensive language dataset (gold). In: Proceedings of the 2021 Conference on Indian Languages Processing. 2021. p. 1–10.

[40] Naik S, P S. Kannada offensive speech dataset (KOSD). Int J Lang Technol 2021;8:98–108.

[41] Rao V, K M. Indic offensive content dataset. In: Proceedings of the 2021 Workshop on Indian Language Processing. 2021. p. 30–40.

[42] Ge L, Moh TS. Improving text classification with word embedding. In: 2017 IEEE International Conference on Big Data (Big Data). IEEE; 2017. p. 1796–1805.

[43] Dadure P, Pakray P, Bandyopadhyay S. BERT-based embedding model for formula retrieval. In: CLEF Working Notes. 2021. p. 36–46.

[44] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–297.

[45] Russell S, Norvig P. Artificial intelligence: A modern approach. 3rd ed. Upper Saddle River (NJ): Pearson Education.

[46] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[47] Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. Hoboken (NJ): Wiley; 2000.

[48] Cover T, Hart P. Nearest-neighbor pattern classification. IEEE Trans Inf Theory 1967;13:21–27.

[49] Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106.

[50] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;29:1189–1232.

[51] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55:119–139.

[52] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26. 2013. p. 3111–3119.
[53] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv [Preprint]. 2019:1810.04805.
[54] Haykin S. Neural networks: A comprehensive foundation. 2nd ed. Upper Saddle River (NJ): Prentice Hall; 1998.
[55] Sarthak, Shukla S, Arya KV. Detecting hostile posts using relational graph convolutional network. arXiv [Preprint]. 2021:2101.03485. doi: 10.48550/arXiv.2101.03485.
[56] Breiman L. Bagging predictors. Mach Learn 1996;24:123–140.
[57] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55:119–139.
[58] Wolpert DH. Stacked generalization. Neural Netw 1992;5:241–259.
[59] Kuncheva LI. Combining pattern classifiers: Methods and algorithms. USA: Wiley-Interscience; 2004.
[60] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–1359.
[61] Hospedales TM, Antoniou A, Micaelli P, Storkey A. Meta-learning in neural networks: A survey. IEEE Trans Pattern Anal Mach Intell 2021;44:1–20.
[62] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. p. 6000–6010.
[63] Powers DMW. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2011;2:37–63.
[64] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–874.
[65] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45:427–437.
[66] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–1284.
[67] Waseem Z, Hovy D. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop. 2016. p. 88–93.
[68] Eisenstein J. What to do about bad language on the Internet. In: Proceedings of NAACL-HLT. 2013. p. 359–369.
[69] Alm CO. Subjective natural language problems: Motivations, applications, characterizations, and implications. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011. p. 107–112.
[70] Bender EM, Friedman B. Data statements for NLP: Toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist 2019;7:587–604.
[71] Ruder S, Howard J, Ruder S. Universal language model fine-tuning for text classification. In: Proceedings of ACL. 2018.
[72] Salem M, Shawky D. A hierarchical framework for multi-aspect evaluation of classifiers in cyberbullying detection. Int J Comput Appl 2010;10:10–15.
[73] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–357.
[74] Conneau A, Lample G, Ranzato M, Denoyer L, Jégou H. Word translation without parallel data. In: Proceedings of ICLR. 2018.
[75] Pavlopoulos J, Sorensen J, Dixon L, Thain N, Androutsopoulos I. Toxicity detection: Does context matter? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2020.
[76] Binns R. Fairness in machine learning: Lessons from political philosophy. In: Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency. 2018. p. 149–159.
[77] Veale M, Binns R, Edwards L. Algorithms that remember: Model inversion attacks and data protection law. Philos Trans A Math Phys Eng Sci 2018;376:20180083.
[78] Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. Predicting the type and target of offensive posts in social media. In: Proceedings of NAACL. 2019. p. 1415–1420.
[79] Sun C, Qiu X, Xu Y. How to fine-tune BERT for text classification? In: Proceedings of China National Conference on Chinese Computational Linguistics. 2019. p. 194–206.
[80] Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, et al. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). arXiv [Preprint]. 2020:2006.07235.