Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 7 (2025)

Automated Threat Detection in Social Media: A Review of Advanced Computational Techniques

[1]Muhammad Noman Saleem, [2*]Muhammad Sabir, [3]Mubasher Malik, [4]Talha Farooq Khan, [5]Abdulrehman Arif

Article Details

Keywords: Threat Detection, Social Media, Hate Speech, Cyberbullying, NLP, Machine Learning, Deep Learning, Hybrid Models, Low-Resource Languages, BERT, Explainable

Muhammad Noman Saleem
Department of Computer Science, University of Southern Punjab, Multan.
Nm4079013@gmail.com
Muhammad Sabir
Department of Computer Science, University of Southern Punjab, Multan. Corresponding Author
Email: muhammadsabir@usp.com.pk

Mubasher Malik
Department of Computer Science, University of Southern Punjab, Multan.
mubasher@usp.com.pk

Talha Farooq Khan
Department of Computer Science, University of Southern Punjab, Multan.
talhafarooqkhan@gmail.com

Abdulrehman Arif
Department of Computer Science, University of Southern Punjab, Multan.
khanabdulrehman026@gmail.com

ABSTRACT

The widespread use of social media platforms has significantly amplified the volume and visibility of harmful digital content, including cyberbullying, hate speech, and threatening text. These threats not only endanger mental health and emotional well-being but also compromise public safety and digital harmony. This review paper provides an in-depth analysis of recent advancements in automated threat detection, focusing on the integration of hybrid Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) techniques. The study examines a variety of models from traditional classifiers like SVM and Logistic Regression to modern transformer-based models such as BERT, RoBERTa, DistilBERT, and MuRIL. It highlights challenges in multilingual and low-resource language contexts, emphasizes the value of hyperparameter tuning and feature optimization, and explores methods for real-time deployment and model explainability. By synthesizing literature across multiple datasets, languages, and threat types, this review aims to guide future research and development of intelligent, ethical, and scalable systems for automated content moderation.

DOI: Availability

INTRODUCTION

The advent of social media platforms has revolutionized global communication, enabling individual to share ideas, express opinions, and build communities across borders. Platforms like Twitter, Facebook, YouTube, and Instagram have become primary tools for social interaction, activism, marketing, and political discourse (Mohbey et al., 2025). However, this increased connectivity has also given rise to a darker side of digital interaction namely, the proliferation of cyber threats such as hate speech, cyberbullying, abusive language, and threatening messages (Almufareh et al., 2025). These harmful behaviors pose significant risks to users' mental health, emotional well-being, and even physical safety (Mohbey et al., 2025).

In recent years, the frequency and severity of threatening content on social media have grown alarmingly. Research highlights the adverse psychological effects on victims, ranging from anxiety and depression to suicidal ideation (Pamila & Kannan, 2025). Threatening content not only affects individuals but also contributes to the spread of fear, misinformation, and social polarization. Manual content moderation is neither scalable nor timely, especially given the vast and fast-paced nature of social media data. This has created a pressing demand for intelligent, automated systems capable of identifying and mitigating harmful online behavior (Shrestha & Dave, 2025).

To address these challenges, the research community has increasingly turned to advanced computational techniques such as Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) (Raza et al., 2024). These approaches offer promising solutions for the automated detection and classification of threatening and abusive content. Modern techniques leverage a variety of algorithms from traditional models like Support Vector Machines (SVM) and Decision Trees, to cutting-edge deep learning architectures such as Bidirectional LSTMs, CNNs, and transformer-based models including BERT, RoBERTa, and DistilBERT (Mohbey et al., 2025).

Moreover, recent advancements emphasize hybrid modeling approaches integrating the strengths of both ML and DL models to achieve higher accuracy, robustness, and adaptability in dynamic digital environments (Alabdulwahab et al., 2023). Hyperparameter tuning further enhances the performance of these systems, making them suitable for real-time applications. Feature extraction techniques such as TF-IDF, Word2Vec, GloVe, and FastText play a crucial role in transforming raw text into meaningful representations for model training. The connectivity of the modern world is greater than ever before and billions of individuals are connected through the internet (Malik, Nawaz, et al., 2024). The interconnection of the people using through the social media made them closer like as a distance of pressing few finger tips (E. A. M. Journal, 2024).

People can express their thoughts, likeness, unlikeness, emotions and feelings about any concerning topic over social media (Moshfegh & Lind, 2023).

Thus, social networking sites, such as Facebook and twitter have become basic necessities for communication, and conveying opinions and even initiating social causes such as protests, politics, governance, and the economy and practically all areas of business are affected (Moshfegh & Lind, 2023). However, these platforms also present some drawbacks primarily where most interactions include use of abusive and threatening language. (Adam et al., 2023)

One challenge that has been observed today is the increase in the threat making messages or comments on the social media. These threats, which mostly imply harm to people is not only irritating but it is demoralize an individual's social, mental and emotional health (Saddozai et al., 2022). From the legal and ethical perspective, this kind of content contradicts with principles of safe and respect communication and cause harmful for the world. (Mehmood et al., 2022)

Thus, the objective of this study is to develop an intelligent model for detect threatening text in social media texts. The textual content of the posts shared on social media will be scanned by our model, and the text that can cause harm will be duly noted and addressed (Moshfegh & Lind, 2023). Old approach of moderating content where the content has to be reviewed manually with social media content. In response to this challenge, a better and faster solution is to seek automatic approaches and tools for realistic identification and classification of threatening text (Arshad & Shahzad, 2024).

In the end, this study will seek to apply different methods for real-time threat detection to overcome these challenges. The emphasis is laid on building large-scale learning models suitable for efficient and fast data processing with focus on the multilingual and contextual aspects of data processing (Malik, Younas, et al., 2024). Since the modern social media environment is relative to constant change and since the traditional methods of threat detection are no longer efficient, this study aims at developing and testing a threat detection system that integrates modern technologies that are for instance, natural language processing, computational linguistics and machine learning. This work will also focus on low resource languages to increase effectiveness and efficiency of the solution (Saddozai et al., 2022).

An emerging challenge lies in addressing multilingualism and low-resource languages (Akhter et al., 2023). Many studies have focused on widely spoken languages like English, but there is a growing body of research targeting underrepresented languages such as Urdu, Bangla, Hausa, and

Swedish, where limited datasets and linguistic tools exist. These efforts are crucial for building inclusive and effective threat detection systems.

This review aims to consolidate and analyze existing literature on automated threat detection in social media, focusing on the use of hybrid, hyperparameter optimized models. The paper presents a comparative overview of various techniques, datasets, evaluation metrics, and real-time implementation strategies. It highlights current gaps, emerging trends, and future directions in the field. By synthesizing state-of-the-art methodologies, this review serves as a foundational resource for researchers and practitioners working toward safer, more resilient online communities.

LITERATURE REVIEW

The literature survey begins by outline five fundamental facets within the domain of threatening text: domain of threatening text are cybersecurity, Natural language processing (NLP), Information security, Artificial intelligence (AI) and Machine learning (ML), and Deep earning (DL). These constituent elements represent critical dimensions of research and inquiry surrounding the complex issue of threatening text detection. The following sections will provide an in-depth exploration of each of these components, explaining the existing body of knowledge, and discerning prevalent trends and developments therein. Through this survey, the research aims to offer a comprehensive overview of the extant literature in the field, thereby contributing to the broader understanding and effective mitigation of threatening text detection.

The detection of threatening text can be classified by type of threat, detection methods and context of the text. The main categories are facilitated by direct violence, passive violence, and non-violence, some detection methods based on machine learning, and NLP.

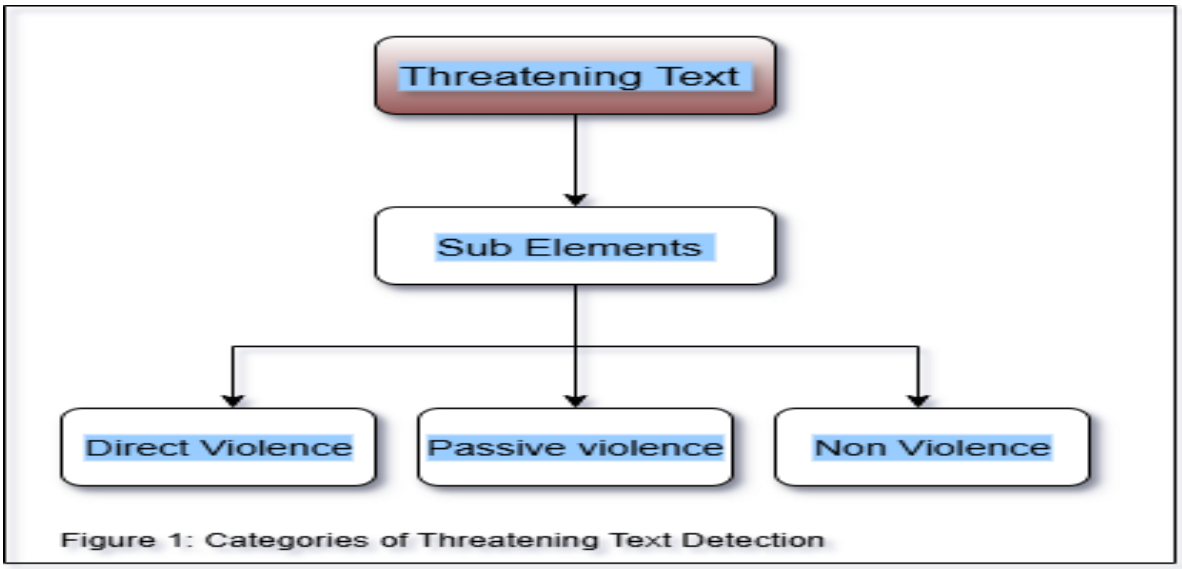

Figure 1: Categories of Threatening Text Detection

DIRECT VIOLENCE

Direct and open violence directed to a person, organization, or group. This is in form of physical threats or death threats, or plans of harm. Examples are I will murder you, We will come and attack you, blow up the building tomorrow. Plenty of the verbs of physical harm, an identified recipient, violence intentions, and Immediate risk producing high priority in detection systems.

PASSIVE VIOLENCE

Passive violence (indirect or implied) threats that emotional abuse, intimidation, or coercion, or incitement to violence but without explicit statements of intent. For example They will get what they deserve, someone ought to give him a lesson, you will rue this etc. It does not contain the expressly worded speech but rather includes veiled or suggestive language, which expresses or incites violence. Tend to be hard to notice because of the attribution to subtlety and the contextual dependency.

NON VIOLENCE

Words that might be viewed as combative or incorporating held views but are not of any threat to anyone. It contains sarcasm, criticism, disagreement or emotional language. Examples: I despise such a thought, You are wrong, this is really silly. Although violent, these statements are not affiliated with displaying violent intent or threat. It's Significant minimizing the occurrence of false positives of automated detection systems.

TABLE 1: A COMPREHENSIVE OVERVIEW OF THE DIVERSE ML, DL, HYBRID TECHNIQUES, AND ARCHITECTURE FOR THREATENING TEXT DETECTION.

Paper	Technique			Architecture	
Reference	ML	DL	Hybrid	Models	Hybrid Model
(Cuzzocrea et al., 2025)	×	×	✓	×	LSTM+TLA-NET
(I. Journal & Science, 2025)	✓	×	×	SVM, NB and LR	×
(Vol et al., 2025)	×	✓	×	Bi-LSTM, NN, and RNN	×
(Abdulfattah & Abdulrahman, 2025)	×	×	✓	×	RF, NB+CNN, Bi-LSTM
(Prama et al., 2025)	×	✓	×	LSTM,CNN	×
(Potla, 2025)	×	×	✓	×	LSTM-3D CNN
(Anand et al., 2025)	×	✓	×	SVM, LR	×
(Akter & Chowdhury, 2025)	×	×	✓	×	SVM,LR,DT+ LSTM
(Daisy, 2025)	✓	×	×	LR,NB,NN	×
(Pranith et al., 2025)	✓	×	×	SVM,NB,LR	×
(Shrestha & Dave, 2025)	✓	✓	×	Optical Character Recognition (OCR), CNN,LSTM,BLSTM	×
(Almufareh et al., 2025)	×	×	✓	×	SVM's+CNN's
(Sabri & Abdullah, 2025)	✓	×	×	SVM, LR, KNN,GNB	×
(Mabel & Idowu, 2025)	✓	✓	×	BERT,GPT,GNN	×

(Pamila & Kannan, 2025)	×	×	✓	×	LSTM+CNN
(Arshad & Shahzad, 2024)	×	✓	×	XLM Roberta,FinBERT	×
(Malik, Nawaz, et al., 2024)	×	✓	×	Urdu-RoBERTa, Urdu-DistilBERT	×
(Khan et al., 2024)	×	×	✓	×	Urdu RoBERTa, Urdu BERT+CNNs
(Malik, Younas, et al., 2024)	×	✓	×	BERT	×
(Wang, 2023)	×	✓	×	DistilBERT	×
(Rehan et al., 2023)	×	✓	×	RoBERTa, MuRIL	×
(Adam et al., 2023)	×	×	✓	×	BERT+CNN+SVM
(Malik et al., 2023)	×	✓	×	Urdu-BERT+BERT64	×
(Moshfegh & Lind, 2023)	×	✓	×	BERT	×
(Bansod, 2023)	×	×	✓	×	MuRIL embeddings+BERT
(Romero & Nguyen, 2022)	×	×	✓	×	BiLSTM+CNN, Naive Bayes
(Saddozai et al., 2022)	×	✓	×	BiLSTM	×
(Hegde & Shashirekha, 2022)	×	×	✓	×	SVM+DistilBERT
(Mehmood et al., 2022)	×	×	✓	×	mDistilBERT+SVM
(Foundation,	×	×	✓	×	mDistilBERT+SVM

2021)

The theme of cyberbullying in the social media and its detrimental impact on mental health of individuals. It suggests a framework known as trustable LSTM-autoencoder network (TLA-NET) to identify cyberbullying (Cuzzocrea et al., 2025). Optimize the efficiency of classification or categorization of cyberbullying by applying Machine Learning concepts to Modified Term Frequency and Inverse Document Frequency (MTF-IDF). In the study, a hybrid model of deep learning (DL) and machine learning (ML) techniques was used to detect cyber bullying activities on Twitter platform (I. Journal & Science, 2025). The rise of malicious activities in instant chat messengers, including cyberbullying and phishing. It highlights the gap in real-time monitoring systems for detecting suspicious activities across platforms (Vol et al., 2025). The study enhances cyber threat detection in tweets using diffusion models and convolutional neural networks (CNNs). It evaluates the effectiveness of DDPM-v4 and uViT techniques on datasets from GitHub and the X API (Abdulfattah & Abdulrahman, 2025). The paper employs a multi-modal deep learning approach integrating text, image, and video processing for cyber threat detection. It utilizes Transformer-based architectures like BERT for text classification (Potla, 2025).

The growing misuse of digital and social media in the twenty-first century, impacting individuals' lives significantly. The proposed system for automatic cyberbullying detection uses supervised machine learning to identify harmful intent and abusive language (Perera & Fernando, 2024). Deep learning models such as CNN, LSTM, and BLSTM were also referenced in related work. (Almomani et al., 2024) It proposes a hybrid approach combining deep learning and traditional machine learning for improved detection performance. The study utilized a dataset containing over 3000 images for cyberbullying detection on social media platforms.

The increasing use of social media by terrorist organizations for spreading beliefs and recruiting members. It highlights the role of Social Media Intelligence (SOCMINT) in monitoring suspicious content related to terrorism. The paper proposes an AI-based Threat Intelligence tool utilizing Deep Learning techniques for OSINT investigations. It employs big data analytical tools to provide real-time insights for organizations (Biagio et al., 2024).

The study utilized machine learning algorithms including Support Vector Machines, Logistic Regression, Naive Bayes, Decision Tree, and K-nearest neighbors for threat detection(Raza et al., 2024). The article (Arshad & Shahzad, 2024) presents technologically advanced world of easy exchange of materials and efforts at connecting with one another, user

meets new challenges on a daily basis. The spread of hate speech and other negativity it is associated with a major problem as per this case. XLM roberta base, FinBERT and fine tuning of such method are applied in this article.

(Moshfegh & Lind, 2023) This paper focuses on the identification of violent threats in Swedish social media. The paper recognizes the difficulties in manual approach caused by the data volume and the lack of threat-oriented low-resource language collections like Swedish. (Bansod, 2023) The paper express the issue of detecting hate speech in Indian languages as having various challenges closely connected to multilingualism while stressing the importance of proper approaches. (Nasim & Falzon, 2022) The thesis explores various aspects of Natural Language Processing (NLP) to analyze social media text for malicious intent. (Mehmood et al., 2022) This paper seeks to solve the low-resource Urdu language classification issue of identifying threatening content on social media platforms like as Twitter, which may un-settling users. (Sheoran, 2021) The paper discusses the rapid increase of social networking sites and the vast amounts of data generated from them. The primary aim is to detect threats in social media networks using Big Data analytics. (Shishira & Patil, 2021) The paper addresses human trafficking as a global issue affecting millions, particularly children and teens under 14 years old. It emphasizes the role of social media in facilitating illicit activities and the need for real-time detection of suspicious communications.

TABLE 2 : A COMPARATIVE ANALYSIS OF MACHINE AND DEEP LEARNING APPROACHES WITH PERFORMANCE METRICS AND DATASET OVERVIEW

Paper Reference	Algorithms	Performance Matrix				Dataset Used
		Accuracy	Precision	Recall	F1-Score	
(Cuzzocrea et al., 2025)	LSTM+TLA-NET	98.62	98.61	98.71	98.68	TRAC-2 dataset
(I. Journal & Science, 2025)	SVM Naïve bayes MTF-IDF, classifier, Adaboost	83.62 71.25 83.82 82.80	×	×	×	Generated by tweets API v2
(Vol et al., 2025)	LSTM, NN, and RNN	99	99	99	99	×

(Abdulfattah & Abdulrahman, 2025)	Random forest, Naïve bayes, CNN, and Bi-LSTM	81.80	80.49	82.16	81.31	GitHub dataset Contain 21,368 tweets
(Prama et al., 2025)	CNN, LSTM, Logistic regression(LR)	98	97	98	97	BullyBlocker dataset
(Potla, 2025)	LSTM-3D, CNN, BERT, ResNet50	92.4	90.8	89.3	90	large-scale dataset with 50,000 posts
(Akter & Chowdhury, 2025)	SVM, Logistic Regression	90.80	88.60	91.40	89.90	Dataset consists of 10,254 comments
(Daisy, 2025)	Naïve baiyes, Logistic Regression, Neural Network	91.9	82.7	81.1	81.5	Twitter API v2
(Shrestha & Dave, 2025)	Optical Character Recognition (OCR), SVM, and LSTM	96	×	×	×	Twitter API v2
(Almufareh et al., 2025)	SVM's, CNN's	95.35	×	×	95	×
(Sabri & Abdullah, 2025)	SVM, KNN, and LR	96	96	95	95	Iraqi Facebook Comments Dataset (IFCD)
(Mabel & Idowu, 2025)	BERT, GPT, and SVM	×	×	×	×	×
(Pamila &	Word2vec	86	×	×	×	5000 tweets

Kannan, 2025)	embedding and LSTM+CNN					
(Arshad & Shahzad, 2024)	XLM Roberta, FinBERT, and BERT-Hinglish,	95	96	95	95 AUROC: 99%	HateInsight dataset
(Malik, Nawaz, et al., 2024)	Urdu-RoBERTa and Urdu-DistilBERT	86.58	×	×	86.52	×
(Khan et al., 2024)	1D-CNN with word unigram model	89.84	×	×	89.80	Dataset contain 4808 tweets
(Malik, Younas, et al., 2024)	Fine-tuned (BERT) model	88 95.12	×	×	×	CrisisMMD
(Wang, 2023)	Graph attention transformer model	×	×	×	84.61	Multilingual offensive language
(Rehan et al., 2023)	RoBERTa and MuRIL and Multi-lingual threatening text detection (MTCD) system	87.17	×	×	84.51	Bilingual Urdu and English datasets
(Adam et al., 2023)	SVM Random forest Decision tree	89	95	89	87	HOC (Hausa Offensive Content) and

	CNN					HTC (Hausa Threatening Content)
(Malik et al., 2023)	fine-tuned Urdu-BERT	87.5	×	×	87.8	Manually generated
(Moshfegh & Lind, 2023)	BERT	×	60	×	×	Dataset contains 6040 posts
(Bansod, 2023)	fastText, GloVe, DistilBERT, and MuRIL	×	39	84	63	RU-HSD-30K dataset
(Romero & Nguyen, 2022)	BiLSTM CNN LM NB SVM	×	×	×	93% AUC: 84.7%	×
(Saddozai et al., 2022)	BiLSTM	82	82	82	82	Generated byTwitter API2
(Hegde & Shashirekha, 2022)	An Extra tree (ET), and Base Naive Bayes (BNB), meta-leearner logistic regression (LR)	74.01	70.84	75.65	73.99	Dataset consists of 3564 tweets
(Mehmood et al., 2022)	BiLSTM CNN	×	×	×	Score 93% AUC: 84.7%	×
(Foundation,	SVM, logistic	80	×	×	×	×

2021)	regression, neural network, m-BERT and RoBERTa	54

TABLE 3: A NUMERICAL REPRESENTED OF THE REVIEW PAPER WHICH CONTAIN HOW MUCH PAPER DOWNLOADED, PAPER SELECTED, PAPER REVIEWED, ALGORITHM USED AND HIGH ACCURACY ALGORITHM.

Serial No	Source/ Database	Numerical Representation				
		Paper downloaded	Paper Selected	Reviewed paper	Algorithm Used	High Accuraccy
1	IEEE Xplore	7	4	4	SVM, LR, DT	98%
2	SpringerLink	23	15	10	BERT, RoBERT, DistilBERT, SVM, DT, LR	92%
3	Google Scholar	30	15	14	Transformer-based Models, Hybrid CNN-LSTM	91%
4	Science – Direct	15	10	9	BERT, Logistic Regression, Ensemble Models	93%
5	ACM Digital Library	9	8	7	Decision Trees, BiLSTM, GloVe Embedding	94%
6	Total papers	84	52	45	SVM, LSTM,	99%

	BERT, CNN, BiLSTM, TF-IDF, etc.

CHALLENGES AND PROBLEMS

In spite of substantial progress in the area of Natural Language Processing (NLP), as well as Machine Learning (ML), the process of identifying threatening material on social media platforms is quite challenging and constantly changing. The social sites produce an extremely large amount of unstructured, multilingual and in many cases ambiguous text based data and the identification of violent or any harmful intentions is an arduous task to achieve with accuracy and efficiency. The chapter lays down the research problems at the centre of the research conducted as a result of literature review.

The lack of adequate datasets for languages like Bangla and Hindi highlights a significant gap in research availability (Cuzzocrea et al., 2025). There is a gap in addressing the nuances of language, such as slang and memes, complicating detection efforts. Future research could explore integrating advanced natural language processing techniques like transformers and BERT (I. Journal & Science, 2025). The study highlights the gap in understanding the dynamics of abusive behavior in online discussions. Addressing abusive behavior is crucial for fostering a safer online environment, indicating a significant research gap (Vol et al., 2025). The paper highlights the need for larger datasets to avoid skewed or overfitted models, indicating a gap in dataset size considerations. It suggests that inflated performance metrics from small datasets do not reflect real-world applications, pointing to a gap in practical applicability (Abdulfattah & Abdulrahman, 2025).

There is a gap in integrating psychological learning theories with Explainable AI models for cyberbullying detection systems. Challenges remain in data imbalance, contextual understanding, and cross-platform generalization in hate speech classification (Prama et al., 2025). The model is vulnerable to evasive cyber threats, necessitating further adversarial training for enhanced robustness. Current multi-modal models face scalability and performance issues in real-time AI-based moderation systems (Potla, 2025). The study highlights the need for continuous updates to models due to the evolving nature of online communication. There is a gap in addressing the nuances of language, such as slang and memes, complicating detection efforts (Anand et al., 2025).

Existing models are often computationally expensive and time-consuming, indicating a

need for more efficient approaches. Errors in scoring systems can lead to missed subtle negative tones, highlighting a gap in detection accuracy (Shrestha & Dave, 2025). There is limited research on extremism and hate speech in the Iraqi dialect, indicating a significant gap in the literature. The datasets available for Iraqi dialect extremism are very poor, necessitating the creation of more comprehensive datasets (Almufareh et al., 2025).

Evolving tactics by attackers necessitate ongoing adaptation of detection methods to maintain effectiveness. Ethical and legal issues surrounding privacy concerns in monitoring private messages need to be addressed (Sabri & Abdullah, 2025).

Problem Area	Description
Detection of Passive/Implied Threats	Subtle language often goes undetected by conventional models
Multilingual and Low-Resource Gaps	Existing models underperform in less common languages
High False Positives	Non-threatening texts are often misclassified as threats
Scalability and Real-Time Issues	Lack of fast, efficient models for real-world social media deployment
Model Interpretability	Deep learning models often lack transparency in decision-making
Ethical and Privacy Issues	Threat detection must balance accuracy with ethical content handling
Dataset Imbalance	Threat examples are rare, making it hard to train balanced and fair models

FUTURE DIRECTION

Online threats are assumed to grow in complexity as social media is increasingly becoming a global method of communication. Although current studies provided considerable progress in identifying cyberbullying, hate speech and threatening material, the field is still full of problems and areas to be studied. The research should be focused on increasing the intelligence, ethics of the systems, scalability, and inclusivity of the systems in the future. This chapter discusses important topics of future research and development in field of threatening text detection.

The study suggests extending efforts in cyberbullying detection systems across multiple

language contexts. Future work could enhance the understanding of aggressive behavior in textual content on social media platforms (Cuzzocrea et al., 2025). Future research should integrate advanced natural language processing techniques like transformers and BERT for improved text classification (I. Journal & Science, 2025). Future research could explore the integration of diffusion techniques with machine learning models for enhanced data analysis (Abdulfattah & Abdulrahman, 2025).

Hybrid models combining deep learning with explainability techniques are crucial for future hate speech detection. Exploring ensemble approaches in NLP can improve hate speech detection effectiveness (Prama et al., 2025). Future research should focus on adversarial training to enhance model robustness against manipulated content. Integrating Generative Adversarial Networks (GANs) can improve model resilience through adversarial examples. Exploring Graph Neural Networks (GNNs) can enhance detection of collaborative cyber threats (Potla, 2025). Future research should integrate advanced natural language processing techniques like transformers and BERT for improved text classification. Expanding datasets to include diverse examples of cyberbullying, including multimedia data, is essential for comprehensive detection (Anand et al., 2025).

Future research should enhance and expand datasets to include diverse instances and contextual factors in cyberbullying detection for Bangla. Exploring transformer-based architectures like BERT or GPT could improve cyberbullying detection systems (Akter & Chowdhury, 2025).

Future research should address the limitations of current AI detection systems in identifying harmful speech accurately. Exploring ethical considerations in automated detection systems is essential for responsible implementation (Daisy, 2025). Integrating multi-modal data sources, including text, images, and audio, can improve detection of cyberbullying across various content formats. Employing NLP models and computer vision techniques can enhance the analysis of textual content and offensive imagery (Pranith et al., 2025). Future research should focus on improving cyberbullying detection systems' accuracy across various platforms, including social media and gaming environments. Investigating the impact of linguistic diversity on detection models can enhance multilingual functionality in cyberbullying detection (Shrestha & Dave, 2025).

Combining text-based sentiment analysis with other modalities like images and videos could further enhance detection performance. Future research should focus on building pre-trained Iraqi vector models on larger datasets for natural language processing applications. There is a need

for a unified and standardized stem for the Iraqi dialect to facilitate research (Almufareh et al., 2025). Future research should address challenges like bias in sentiment analysis and the occurrence of false positives in flagged messages. Investigating evolving tactics used by misinformation spreaders is essential for improving detection methods (Sabri & Abdullah, 2025).

ADVANCEMENT IN MULTILINGUAL AND LOW-RESOURCE NLP

There is dire need to have fair and inclusive detection models that will support different languages and dialects. Future studies are advised to be concentrated on:

- Multilingual pretrained transformers (e.g. XLM-R, mBERT, MuRIL).
- Cross-lingual training and transfer learning
- Zero- and few-shot learning on low-resource languages

REAL-TIME, LIGHTWEIGHT, AND SCALABLE SYSTEMS

For the real time Future models are to be:

- Computationally efficient
- Low latency-optimized hyperparameter
- It can be implemented on a real-time basis

Live moderation of the threats on a large scale could be discussed with the help of Edge AI and stream processing frameworks such as Apache Kafka or Flink.

(XAI) INCORPORATION OF EXPLAINABLE AI (XAI)

AI-driven systems are in need of more transparency and trust. The next studies must incorporate explainable methods like:

- SHAP (SHapley Additive Explanations)
- LIME (Local Interpretable Model-Agnostic Explanations)
- Visualizing attention in transformers

These techniques will render the model prediction interpretable to the users and the moderators and this way there will be an accountability in sensitive setups.

FUSION OF MULTI-MODAL DATA FOR THREAT DETECTION

Sometimes the entire meaning of a post might not be put through text. The future threat detection systems will be capable of integrating text, pictures, videos, and metadata to be able to analyze it more thoroughly. As an illustration, a meme that involves violent pictures and aggressive texts is worse than either one. Robustness and accuracy can be enhanced greatly by multi-modal models.

CREATION OF BENCHMARK DATASETS AND ANNOTATION TOOLS

In the future work it would also be worth prioritizing:

- Creation of annotated, multi-language, and morally-neutral data (sets)
- Semi-privacy preserving annotation methods by crowdsourcing
- Data generation (e.g., with the use of LLMs) to represent underrepresented threat categories

CONCLUSION

Threat detection on social media is an urgent and complex problem that demands sophisticated technological solutions. The review reveals that while significant strides have been made using ML, DL, and transformer-based NLP models, major challenges remain particularly in detecting implied threats, handling multilingual content, reducing false positives, and ensuring ethical implementation. Hybrid models that combine the strengths of deep contextual embeddings and traditional classifiers demonstrate improved accuracy and robustness, especially in real-time systems.

Furthermore, low-resource languages continue to present a major research gap, with the need for more inclusive datasets and transfer learning techniques. Future systems must integrate explainable AI for transparency, support multi-modal data for richer context, and be lightweight enough for deployment on real-time platforms. Ethical considerations such as privacy, fairness, and legal compliance must also be embedded in system design.

This paper contributes to the growing body of work that seeks to make online spaces safer through AI-driven content moderation. It serves as a foundational reference for researchers and developers aiming to build the next generation of intelligent, accountable, and inclusive threat detection systems for digital environments.

REFERENCES

Abdulfattah, A., & Abdulrahman, M. (2025). *RIT Digital Institutional Repository Enhancing Cyber Threat Detection in Tweets Using Diffusion Models and Convolutional Neural Networks.*

Adam, F. M., Zandam, A. Y., & Inuwa-Dutse, I. (2023). *Machine Learning for Identifying Harmful behaviour on social media.* http://arxiv.org/abs/2311.10541

Ahmed, M. T., Antar, A. H., Rahman, M., Islam, A. Z. M. T., Das, D., & Rashed, M. G. (2023). Social Media Cyberbullying Detection on Political Violence from Bangla Texts Using Machine Learning Algorithm. *Journal of Intelligent Learning Systems and Applications, 15*(04),

108–122. https://doi.org/10.4236/jilsa.2023.154008

Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2023). A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Natural Language Processing Journal*, *4*(July), 100027. https://doi.org/10.1016/j.nlp.2023.100027

Akter, F., & Chowdhury, R. R. (2025). *Cyberbullying Detection on Social Media Platforms Utilizing Different Cyberbullying Detection on Social Media Platforms Utilizing Different Machine Learning Approaches*. *January*. https://doi.org/10.5120/ijca2025924395

Alabdulwahab, A., Haq, M. A., & Alshehri, M. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. *International Journal of Advanced Computer Science and Applications*, *14*(10), 424–432. https://doi.org/10.14569/IJACSA.2023.0141045

Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic Cyberbullying Tweets Using Machine Learning. *Machine Learning and Knowledge Extraction*, *5*(1), 29–42. https://doi.org/10.3390/make5010003

AlGhamdi, M. A., & Khan, M. A. (2020). Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages. *Arabian Journal for Science and Engineering*, *45*(8), 6021–6032. https://doi.org/10.1007/s13369-020-04447-0

Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image cyberbullying detection and recognition using transfer deep machine learning. *International Journal of Cognitive Computing in Engineering*, *5*(December 2023), 14–26. https://doi.org/10.1016/j.ijcce.2023.11.002

Almufareh, M. F., Jhanjhi, N., Humayun, M., Alwakid, G. N., Javed, D., & Almuayqil, S. N. (2025). Integrating Sentiment Analysis with Machine Learning for Cyberbullying Detection on Social Media. *IEEE Access*, *PP*(January), 1. https://doi.org/10.1109/ACCESS.2025.3558843

Alqahtani, A. F., & Ilyas, M. (2024). An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying. *Machine Learning and Knowledge Extraction*, *6*(1), 156–170. https://doi.org/10.3390/make6010009

Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A., & Gelbukh, A. (2021). Threatening Language Detection and Target Identification in Urdu Tweets. *IEEE Access*, *9*(September), 128302–128313. https://doi.org/10.1109/ACCESS.2021.3112500

Anand, A. K., Mahto, R. K., & Prasad, A. (2025). *Analysis of Cyberbullying Behaviors Using Machine Learning : A Study on Text Classification Análisis de los comportamientos de ciberacoso mediante*

aprendizaje automático : Un estudio sobre clasificación de textos. https://doi.org/10.62486/latia2025126

Arshad, M. U., & Shahzad, W. (2024). Understanding hate speech: the HateInsights dataset and model interpretability. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/PEERJ-CS.2372

Bansod, P. P. (2023). *Hate Speech Detection in Hindi.* https://scholarworks.sjsu.edu/etd_projects/1265

Biagio, M. S., Simoncini, S., La Mattina, E., & Morreale, V. (2024). MARPLE: A Framework for Social Media Threat Intelligence. *International Conference on Artificial Intelligence, Computer, Data Sciences, and Applications, ACDSA 2024, February*, 1–2. https://doi.org/10.1109/ACDSA59508.2024.10467738

Cuzzocrea, A., Akter, M. S., Shahriar, H., & Garcia Bringas, P. (2025). Cyberbullying Detection, Prevention, and Analysis on Social Media via Trustable LSTM-Autoencoder Networks over Synthetic Data: The TLA-NET Approach †. *Future Internet*, *17*(2). https://doi.org/10.3390/fi17020084

Daisy, E. (2025). *AI-Powered Social Media Monitoring : Leveraging Natural Language Processing for Real-Time Cyberbullying Detection on Twitter Author : Evelyn Daisy Date : April 2025. April.*

Fakhouri, H. N., Alhadidi, B., Omar, K., Makhadmeh, S. N., Hamad, F., & Halalsheh, N. Z. (2024). AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response. *2nd International Conference on Cyber Resilience, ICCR 2024.* https://doi.org/10.1109/ICCR61006.2024.10533010

Fkih, F., & Al-Turaif, G. (2023). Threat Modelling and Detection Using Semantic Network for Improving Social Media Safety. *International Journal of Computer Network and Information Security*, *15*(1), 39–53. https://doi.org/10.5815/ijcnis.2023.01.04

Fonseca Abreu, J. V., Ghedini Ralha, C., & Costa Gondim, J. J. (2021). *A Multi-agent Approach for Online Twitter Bot Detection. June.* https://doi.org/10.18239/jornadas_2021.34.03

Foundation, J. (2021). *AWARENESS OF DIGITAL ABUSE AMONG COLLEGE John Foundation Journal of EduSpark. 3*(3), 29–35.

Granizo, S. L., Caraguay, A. L. V., Lopez, L. I. B., & Hernandez-Alvarez, M. (2020). Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. *IEEE Access*, *8*, 44534–44546. https://doi.org/10.1109/ACCESS.2020.2976530

Hegde, A., & Shashirekha, H. L. (2022). Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets. *CEUR Workshop Proceedings*, *3395*, 256–265.

Journal, E. A. M. (2024). *CYBERBULLYING : ITS SOCIAL AND ACADEMIC EFFECTS TO UNDERGRADUATE STUDENTS IN THE PROVINCE OF CAPIZ Cyberbullying : Its Social and Academic Effects to Undergraduate Students in the Province of Capiz*. *5*, 715–724. https://doi.org/10.5281/zenodo.13824772

Journal, I., & Science, O. F. (2025). *Detecting Cyberbullying in Social Media : An NLP-Based Classification*. 380–389.

Khan, M. S., Malik, M. S. I., & Nadeem, A. (2024). Detection of violence incitation expressions in Urdu tweets using convolutional neural network. *Expert Systems with Applications*, *245*(January), 123174. https://doi.org/10.1016/j.eswa.2024.123174

Kour, I., & Technology, I. (2024). *Phrase-wise Dataset Development on Cyberbullying : A social media perspective*.

Mabel, E., & Idowu, M. (2025). *Sentiment Analysis in Social Media : Detecting Misinformation and Cyber Threats*. *February*.

Malik, M. S. I., Cheema, U., & Ignatov, D. I. (2023). Contextual Embeddings based on Fine-tuned Urdu-BERT for Urdu threatening content and target identification. *Journal of King Saud University - Computer and Information Sciences*, *35*(7), 101606. https://doi.org/10.1016/j.jksuci.2023.101606

Malik, M. S. I., Nawaz, A., & Jamjoom, M. M. (2024). Hate Speech and Target Community Detection in Nastaliq Urdu Using Transfer Learning Techniques. *IEEE Access*, *12*(July), 116875–116890. https://doi.org/10.1109/ACCESS.2024.3444188

Malik, M. S. I., Younas, M. Z., Jamjoom, M. M., & Ignatov, D. I. (2024). Categorization of tweets for damages: infrastructure and human damage assessment using fine-tuned BERT model. *PeerJ Computer Science*, *10*, 1–27. https://doi.org/10.7717/peerj-cs.1859

Mehmood, A., Farooq, M. S., Naseem, A., Rustam, F., Villar, M. G., Rodríguez, C. L., & Ashraf, I. (2022). Threatening URDU Language Detection from Tweets Using Machine Learning. *Applied Sciences (Switzerland)*, *12*(20). https://doi.org/10.3390/app122010342

Mohbey, K. K., Sterjanov, M., & Margarita, V. (2025). *Hate Speech Identification and Categorization on Social Media Using Bi-LSTM : An Information Science Perspective*. *13*(1), 51–69.

Moshfegh, A., & Lind, K. (2023). *USING MACHINE LEARNING TO IDENTIFY HATE*

SPEECH ON SOCIAL MEDIA.

Murad, S. A., Dahal, A., & Rahimi, N. (n.d.). *Multi-Lingual Cyber Threat Detection in Tweets / X Using ML , DL , and LLM : A Comparative Analysis*. 1–12.

Nasim, M., & Falzon, A. P. G. (2022). *Analysing language use on social media for identifying malicious activities . Pranav Bhandari Supervisors Submitted in partial fulfillment of the requirements for the degree*.

Pamila, M. J., & Kannan, S. (2025). *Improved Detection of Suicidal Tendency from Tweets using Stacked LSTM IMPROVED DETECTION OF SUICIDAL TENDENCY FROM TWEETS USING STACKED LSTM CNN MODEL*. *June 2024*. https://doi.org/10.5281/zenodo.15174126

Perera, A., & Fernando, P. (2024). Cyberbullying Detection System on Social Media Using Supervised Machine Learning. *Procedia Computer Science*, *239*(2023), 506–516. https://doi.org/10.1016/j.procs.2024.06.200

Potla, R. T. (2025). *AI-Powered Threat Detection in Online Communities : A Multi-Modal Deep Learning Approach*. *13*(2), 155–171. https://doi.org/10.4236/jcc.2025.132010

Prama, T. T., Amrin, J. F., Anwar, M. M., & Sarker, I. H. (2025). *AI Enabled User-Specific Cyberbullying Severity Detection with Explainability*. 1–25. http://arxiv.org/abs/2503.10650

Pranith, G., Krishna, B. Y., Krupa, D. L., & Venkatesh, G. A. (2025). *International Journal of Advanced Research in Education and TechnologY (IJARETY) Machine Learning Solutions for Cyberbullying Detection and Prevention on Social Media*. *12*(2). https://doi.org/10.15680/IJARETY.2025.1202051

Raza, M. O., Meghji, A. F., Mahoto, N. A., Al Reshan, M. S., Abosaq, H. A., Sulaiman, A., & Shaikh, A. (2024). Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data. *International Journal of Computational Intelligence Systems*, *17*(1). https://doi.org/10.1007/s44196-024-00580-y

Rehan, M., Malik, M. S. I., & Jamjoom, M. M. (2023). Fine-Tuning Transformer Models Using Transfer Learning for Multilingual Threatening Text Identification. *IEEE Access*, *11*(August), 106503–106515. https://doi.org/10.1109/ACCESS.2023.3320062

Romero, P. M., & Nguyen, D. (2022). *Classifying Legally Actionable Threats Using Language Models Noud Jan de Rijk (5670721). 5670721*.

Sabri, R. F., & Abdullah, N. A. Z. (2025). Extremism Detection in the Iraqi Dialect Based on Machine Learning. *Iraqi Journal of Science*, *66*(2), 876–889.

https://doi.org/10.24996/ijs.2025.66.2.25

Saddozai, F. K., Ahmad, H., Asghar, M. U., & Khan, A. (2022). *Hate Speech Detection from Urdu language Tweets using Deep Learning Technique Bidirectional Encoder Representations from Transformers*. *14*(1), 523–537.

Sheoran, S. K. (2021). *THE EFFICIENT MANAGEMENT FOR MALICIOUS USER DETECTION IN BIG S . info DETECTION IN BIG DATA Threats Detection using Big Data Analytics*. *December*.

Shishira, S. S., & Patil, J. S. (2021). *DETECTION OF ILLICIT MESSAGES IN TWITTER USING SUPPORT VECTOR MACHINE AND VGG16*. *9*(3), 794–804.

Shrestha, R., & Dave, R. (2025). *Machine Learning for Identifying Harmful Online Behavior : A Cyberbullying Overview*. *13*(1), 26–40. https://doi.org/10.4236/jcc.2025.131003

Vol, F., Journal, F., & Sciences, A. (2025). *Development of an Intelligent Chat Monitoring and Suspicious Activity Detection System S. A. Mogaji*, T.T. Odufuwa, C.A. Afolalu, O.O Faboya, M.V. Ige, and O.D. Idowu Department of Computer Science, Federal University Oye-Ekiti, Ekiti State, Nigeria*. *10*, 0–3.

Wang, G. (2023). MLOffense: Multilingual offensive language detection and target identification on social media using graph attention transformer. *Applied and Computational Engineering*, *21*(1), 36–46. https://doi.org/10.54254/2755-2721/21/20231114