Data Security by Using Machine Learning and Deep Learning

Gaurav Aggarwal[a] and Pooja[b]

[a] Professor, Department of CSE, Jagannath University, Bahadurgarh, Jhajjar (Haryana)

[b] Research Scholar, Department of CSE, Jagannath University, Bahadurgarh, Jhajjar (Haryana)

Abstract- This paper proposes a hybrid architecture that integrates Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs) to detect and mitigate social media threats in real time. The proposed system incorporates preprocessing, multi-layered classification, and automated mitigation modules, and leverages fine-tuned transformer models like BERT, RoBERTa, and GPT-3. Experimental results across multiple datasets show that LLMs outperform traditional models in identifying nuanced threats such as sarcasm, coded hate speech, and misinformation, achieving precision rates above 95%. Furthermore, case studies in cyberbullying, fake news detection, and bot activity highlight the system's real-world applicability. The findings emphasize the need for ethical, explainable, and scalable AI-driven solutions to enhance safety and integrity on social media platforms.

Keywords- Machine Learning, Deep Learning, Large Language Models, Cybersecurity, NLP, Content Moderation.

I. Introduction

In recent years, advancements in Machine Learning (ML), Deep Learning (DL), and more recently, Large Language Models (LLMs) have opened new avenues for building intelligent, adaptive, and scalable security solutions for social media. Traditional ML models such as Support Vector Machines (SVM), Naïve Bayes, and Random Forest have been widely used for spam detection, sentiment analysis, and basic content classification [5]. DL models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), introduced the ability to learn hierarchical or sequential patterns directly from raw text and image data, significantly enhancing detection accuracy in complex scenarios [6]. However, both ML and standard DL techniques fall short in capturing semantic nuance, cross-sentence dependencies, and implicit meaning, which are often critical in identifying hate speech, sarcasm, and misinformation [7].

The advent of transformer-based LLMs, such as BERT [8], GPT-3 [9], RoBERTa [10], and T5 [11], has significantly improved the field of Natural Language Understanding (NLU) by modeling long-range contextual dependencies through attention mechanisms. These models are pre-trained on massive corpora and can be fine-tuned for specific downstream tasks with high accuracy, even in few-shot or zero-shot settings. LLMs are particularly effective at identifying implicit threats, emotive subtext, coded language, and multi-lingual variations, all of which are frequently encountered in harmful or deceptive social media content [12], [13].

This paper proposes a hybrid architecture that integrates ML, DL, and LLM models into a unified pipeline to detect and mitigate a wide range of social media threats in real-time. The primary contributions of this research are as follows:

- A taxonomy of social media threats, categorizing them by modality (text, image, video), intent (malicious vs. ignorant), and impact (individual harm, public misinformation, platform integrity).
- A novel LLM-enhanced classification and mitigation framework, combining contextualized language models (e.g., GPT-3, BERT) with automated moderation strategies such as threat flagging, user feedback loops, and dynamic blacklist generation.
- A comparative performance evaluation of traditional ML models, deep learning architectures (CNN, RNN), and LLMs across multiple publicly available datasets, including hate speech, cyberbullying, and misinformation benchmarks.

The results of this study indicate that LLM-based models not only outperform conventional classifiers in terms of accuracy, recall, and F1-score, but also exhibit superior capabilities in identifying subtle, sarcastic, and coded language, making them ideal candidates for robust, real-time social media moderation.

II. Background and Related Work

Security and moderation challenges on social media platforms have been a significant concern over the past decade. As platforms grow in reach and influence, they become increasingly susceptible to cyber threats such as cyberbullying, misinformation, coordinated disinformation campaigns, hate speech, spam, and bot attacks. To address these threats, a wide array of techniques ranging from rule-based systems to large-scale deep learning models have been explored. This section reviews the evolution of such techniques in the context of secure social media analysis and content moderation.

A. Traditional Approaches

Early research on social media security primarily relied on keyword-based filters and manual moderation processes. These techniques involved static lists of offensive words, profanity lexicons, and regular expression patterns to detect harmful content [14]. Although straightforward, such methods suffered from several limitations: they were easily circumvented by the use of alternative spellings (e.g., leetspeak), resource-intensive, and often failed to capture the nuanced or implicit meanings of harmful messages [3]. Manual moderation, while more accurate, is not scalable and places a considerable burden on human moderators leading to delays and potential psychological stress due to constant exposure to toxic content [15].

B. Machine Learning-Based Systems

To enhance scalability and improve detection accuracy, researchers adopted supervised machine learning algorithms for content classification tasks. Models such as Support Vector Machines (SVM), Random Forests, Decision Trees, and Naïve Bayes classifiers were trained on labeled datasets to detect spam, hate speech, fake accounts, and offensive content [16], [5]. For instance, Waseem and Hovy (2016) demonstrated the use of logistic regression and SVMs for hate speech detection on Twitter [17]. While these models significantly outperformed rule-based approaches, they often required careful feature engineering, including TF-IDF vectors, part-of-speech tagging, and n-gram extraction. However, these traditional ML models lacked the ability to generalize across domains or detect implicit or sarcastic language, thus limiting their effectiveness in real-world applications.

C. Deep Learning Enhancements

With the advent of deep learning, researchers began employing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically learn feature representations from raw text and images. CNNs were effective in detecting local text patterns and visual features in memes or hate images, while RNNs (and their variants such as LSTMs and GRUs) were used to model sequential relationships in language, enabling more coherent interpretation of longer posts and comment threads [18]. For instance, Zhang et al. (2018) applied CNN-LSTM architectures for toxic comment detection and reported notable improvements over ML baselines [19]. Despite these advances, deep learning models still lacked deep contextual understanding and often failed to recognize humor, sarcasm, or subtle forms of misinformation due to their limitations in capturing long-range dependencies and inter-sentential context.

D. Rise of Large Language Models (LLMs)

The introduction of transformer-based models, starting with BERT (Bidirectional Encoder Representations from Transformers), marked a significant breakthrough in natural language understanding (NLU). Unlike RNNs, transformers utilize self-attention mechanisms to capture dependencies between words regardless of their distance, thereby enabling robust modeling of context and semantics. This has led to substantial improvements in classification, summarization, question-answering, and content moderation tasks. Models like GPT-2/GPT-3, RoBERTa, and T5 have shown state-of-the-art performance across a variety of benchmarks, including hate speech detection, rumor detection, and sentiment analysis. [10-13]

Recent studies have demonstrated that LLMs significantly outperform traditional and deep learning models in tasks involving implicit hate speech, sarcasm, and politically sensitive misinformation. For example, Liu et al. (2022) showed that RoBERTa achieved over 93% accuracy in detecting coded racism and offensive content that eluded CNN and SVM models [20]. Furthermore, LLMs can be fine-tuned for domain-specific moderation tasks or adapted in few-shot and zero-shot settings, enabling flexible deployment across multiple platforms with minimal labeled data. They are also capable of modeling user behavior and discourse dynamics, making them highly suitable for identifying coordinated disinformation campaigns and bot activities [21].

Alex Kaplunovich et al. (2024) This study highlights the growing privacy vulnerabilities in social media platforms, driven by the vast user data shared voluntarily through APIs. Utilizing cloud-based serverless architectures, machine learning graph models, and large language models (LLMs), the research demonstrates how personal information such as geolocation, social links, and metadata from uploaded photos can be efficiently harvested, analyzed, and clustered at scale. The use of tools like NoSQL DynamoDB and Generative AI (e.g., RAG) enables accurate inference of user behavior, influencer identification, and hidden network connections.

A key concern raised is the exploitation of photo metadata, which can reveal timestamps, GPS coordinates, and device details transforming smartphones into continuous tracking devices. This aspect adds a critical cybersecurity layer to digital privacy debates. The paper underscores the need for user caution and urges social media platforms to reassess third-party access policies, advocating for stronger data governance, transparency, and ethical AI deployment to mitigate surveillance risks.

TALHA et al. (2024) The proliferation of social media platforms has transformed communication, but it has also given rise to social media bots that can spread misinformation, manipulate public opinion, and compromise the integrity of online discourse. This thesis addresses the critical issue of detecting social media bots on Twitter. Traditional detection methods often fall short due to the evolving nature of these bots and the vast amount of data involved. To overcome these challenges, this research proposes a hybrid ensemble model that combines profile-based and content-based features with advanced natural language processing techniques. This approach captures a wide range of bot behaviors and characteristics, resulting in more accurate and robust detection. The thesis includes an examination of social media platforms and the threats posed by bots, a review of current bot detection methods, an in-depth explanation of the proposed hybrid ensemble methodology, and an experimental evaluation of the methodology's effectiveness compared to leading techniques. The findings demonstrate significant improvements in detection performance, supporting efforts to protect social media environments from harmful automated entities.

III. Taxonomy of Threats on Social Media

A comprehensive and structured classification of threats on social media is a critical prerequisite for developing scalable, intelligent, and context-sensitive security frameworks. Social media threats vary significantly in form, origin, and impact, making it essential to categorize them based on their underlying nature and behavior. In this study, threats are categorized into four primary dimensions: Content-Based, Account-Based, Network Based, and Behavioral-Based. Each category encompasses a unique set of risks and requires different detection mechanisms and mitigation strategies.

A. Content-Based Threats

Content-based threats refer to harmful or malicious user-generated posts, comments, or multimedia that target individuals, communities, or institutions. These include:

- **Hate Speech:** Language that attacks or demeans individuals based on race, religion, gender, or other identity markers. Hate speech is often implicit, relying on euphemisms, codewords, or sarcasm to evade detection [24].
- **Cyberbullying:** Repeated, intentional online harassment or threats directed at individuals, often teenagers or public figures, through mocking, exclusion, or verbal abuse. It frequently appears in private messages or public comment threads [19].
- **Misinformation and Fake News:** The spread of misleading or false information, intentionally or unintentionally, on sensitive topics such as health, politics, or science. Notably, COVID-19 vaccine misinformation is a widely studied example [25].

Traditional ML and keyword-based models struggle to capture the semantic subtlety, emotional tone, and contextual dependencies involved in hate speech and cyberbullying. Moreover, misinformation detection often requires external knowledge or fact verification, which cannot be done using syntactic features alone.

B. Account-Based Threats

Account-based threats involve the manipulation or abuse of social media accounts for malicious purposes. Typical threats include:

- **Bot Detection:** Automated accounts that spread propaganda, amplify misinformation, or manipulate trends using scripted behavior. Bots can be simple rule-based scripts or more advanced AI-generated personas [26].
- **Fake Accounts:** Profiles created with stolen or fabricated identities to deceive users, inflate follower counts, or perform scams.
- **Impersonation:** When a malicious actor mimics a real user or organization to mislead others, often using similar usernames, images, or stolen content.

These threats require modelling of account metadata, network patterns, and posting behavior. Traditional ML methods are often limited due to incomplete feature representations, while LLMs can provide context-sensitive evaluations when combined with behavior logs and user history.

C. Network-Based Threats

Network-based threats leverage the underlying connectivity and communication architecture of social media to propagate harm. Examples include:

- **Phishing:** Malicious attempts to obtain personal information through deceptive messages or links disguised as legitimate sources.
- **Malware Links:** Distribution of URLs or attachments that lead to drive-by downloads, ransomware, or spyware infections.
- **Social Engineering:** Manipulating individuals into revealing confidential data or performing actions (e.g., clicking links, donating money), often through psychological manipulation.

Network-based threats require link analysis, domain reputation checks, and real-time URL classification, often relying on external threat intelligence feeds. LLMs can aid in textual phishing detection by identifying patterns in suspicious messages (e.g., urgent tone, financial language, spoofed branding).

D. Behavioral-Based Threats

Behavioral-based threats focus on detecting patterns of user interaction or platform usage that deviate from typical behavior. These include:

- **Anomalous Activity:** Sudden bursts of messages, login attempts from unusual locations, or massive retweeting within seconds, which may indicate account hijacking or bot activity [27].
- **Coordinated Attacks:** Multiple accounts posting similar messages in sync, often part of disinformation campaigns or political manipulation.
- **Online Radicalization:** Progressive exposure to and sharing of extremist content that leads to ideological alignment or offline actions.

These threats require temporal modeling, graph analysis, and sequence-based user modeling. Behavioral patterns are complex, user-specific, and context-dependent, making them ideal for LLM-augmented anomaly detection systems that integrate temporal embeddings and attention-based behavior modelling.

Table 1: Taxonomy of Threats on Social Media

Category	Example Threats
Content-Based	Hate speech, cyberbullying, misinformation
Account-Based	Bot detection, fake accounts, impersonation
Network-Based	Phishing, malware links, social engineering
Behavioral-Based	Anomalous activity, coordinated attacks, radicalization

The diversity of threats across these four categories illustrates the multi-modal and multi-layered nature of social media risks. As platforms grow in complexity, it becomes increasingly important to use multi-model hybrid systems combining ML, DL, and LLMs to effectively address threats. For instance, while CNNs may detect offensive image content, LLMs like GPT-3 or RoBERTa can understand context in textual threats, and graph neural networks (GNNs) may be used to uncover coordinated network activity. Integrating this taxonomy into system design allows for modular security pipelines, where different models specialize in detecting distinct threat types, collectively improving the robustness of social media moderation systems.

IV. Proposed Architecture

A. System Overview

The proposed architecture is designed as a modular, scalable, and intelligent framework for identifying, classifying, and mitigating social media threats. It leverages traditional ML models, advanced DL architectures, and cutting-edge Large Language Models (LLMs) to offer a multi-layered defense mechanism. The architecture is divided into three major modules:

1) Preprocessing Layer

The preprocessing layer is responsible for cleansing and normalizing the input data before feeding it into the models. This includes:

- **Tokenization:** The raw text is broken into individual tokens (words or subwords) using tools like SpaCy, NLTK, or tokenizer APIs from the HuggingFace library.
- **Noise Removal:** Unnecessary elements such as HTML tags, emojis, URLs, special characters, and stopwords are removed to enhance signal-to-noise ratio.

Embedding Techniques:

- Word2Vec and GloVe are used for lightweight ML models.
- BERT Embeddings are used to retain contextual relationships for transformer-based models.

2) Classification Layer

The classification layer is a hybrid ensemble of three model families:

Machine Learning Models

- **Support Vector Machine (SVM):** Effective for binary classification such as toxic vs. non-toxic content.
- **XGBoost:** Provides gradient-boosted decision trees for high accuracy in multiclass classification tasks.

Deep Learning Models

- **Convolutional Neural Networks (CNN):** Used for detecting visual patterns in text-like sequences (e.g., character n-grams).
- **Long Short-Term Memory Networks (LSTM):** Efficient in capturing temporal patterns and dependencies in sequential user posts or interactions.

Large Language Models (LLMs)

- **GPT (Generative Pre-trained Transformer):** Handles generation and detection tasks like sarcasm or misinformation.
- **RoBERTa (Robustly Optimized BERT):** Ideal for sentence-level classification due to robust pretraining on large-scale data.

3) Mitigation Layer

Once the input is classified, the mitigation layer executes automatic countermeasures based on the classification results:

- **Auto-Moderation:** Flags and hides offensive content automatically using rule-based triggers linked with model confidence scores.
- **User Flagging and Alerting:** Suspicious users (e.g., spammers, bots) are flagged and forwarded to human moderators.
- **Threat Report Generation:** Structured summaries of flagged content, risk level, and potential impact are generated in JSON or PDF formats for audit logs or law enforcement.

B. Model Selection

The choice of models is guided by the nature of the input data, computational efficiency, and task complexity:

- **ML Models (SVM, Logistic Regression, Random Forest, XGBoost):** Suited for structured, labeled datasets with features such as comment length, frequency, user metadata, and sentiment scores.
- **DL Models (CNN, LSTM, BiLSTM, GRU):** These models excel in capturing hierarchical or temporal patterns in sequences of posts, comments, or tweet threads. CNNs are fast for local feature extraction, while LSTMs provide memory for long-range dependencies.
- **LLM Models (BERT, RoBERTa, GPT-2, GPT-3):** LLMs can analyze complex linguistic structures and hidden sentiments such as irony, coded hate speech, and emergent slangs. They are particularly effective in zero-shot or few-shot learning scenarios.

C. Data Sources

To train and evaluate the system, diverse datasets from multiple platforms were curated:

Twitter & Reddit Datasets: Publicly available corpora annotated for:

- Hate speech (e.g., Davidson et al., 2017)
- Cyberbullying (e.g., OLID, Waseem & Hovy, 2016)
- Misinformation & propaganda (e.g., FakeNewsNet, PHEME)

Facebook Comment Dataset: A structured dataset containing thousands of multilingual, user-generated comments labeled as offensive, neutral, or benign. Available via CrowdFlower or Kaggle.

Kaggle Datasets

- Cyberbullying Detection Dataset (Kaggle, 2021)
- Fake News Challenge Dataset
- Toxic Comment Classification Dataset (Jigsaw)

All datasets were preprocessed and augmented using paraphrasing and back-translation to improve generalizability.

D. LLM Integration

The core strength of the system lies in the incorporation and fine-tuning of transformer-based Large Language Models.

Pretrained Models

- **BERT (Bidirectional Encoder Representations from Transformers):** Offers strong performance on classification tasks.
- **RoBERTa:** Optimized variant of BERT with improved training methodology.
- **GPT-3:** Capable of both classification and generation, ideal for threat detection and synthetic content moderation.
- **Fine-tuning Objective:** Each LLM is fine-tuned for binary (e.g., toxic vs. non-toxic) or multiclass (e.g., cyberbullying, hate speech, sarcasm, spam) classification using labeled datasets. Training involves:

- Cross-entropy loss for classification.
- Learning rate warmup and cosine decay.
- Layer-wise learning rate tuning to reduce catastrophic forgetting.

Tools and Frameworks

- **HuggingFace Transformers:** Provides APIs for model loading, tokenization, and fine-tuning.
- **PyTorch & TensorFlow:** For defining custom architectures, implementing attention masking, and distributed training.
- **ONNX Runtime:** For optimized inference and integration with real-time social media monitoring tools.

V. Experimental Setup and Results

This section presents the experimental configuration, evaluation methodology, and comparative analysis of different models used for detecting and mitigating threats on social media platforms.

A. Evaluation Metrics

To evaluate the performance of the classification models, we use five well-established metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Each metric captures a different aspect of classification effectiveness, especially in imbalanced or multi-class scenarios.

Let:

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

The performance metrics are defined as follows:

1. **Accuracy:** Accuracy measures the proportion of correct predictions among the total number of cases evaluated.

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

2. **Precision:** Precision measures how many of the predicted positive classes are actually positive. It is critical for minimizing false alarms in security detection.

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

3. **Recall (Sensitivity):** Recall quantifies the model's ability to identify all relevant instances in the dataset.

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

4. **F1-Score:** The F1-score is the harmonic mean of Precision and Recall, and it balances both false positives and false negatives.

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

5. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** ROC-AUC provides an aggregate measure of performance across all classification thresholds. AUC values range from 0.5 (random guessing) to 1.0 (perfect classification).

B. Performance Comparison

We compared multiple models traditional machine learning (SVM), deep learning (CNN, LSTM), and large language models (BERT, GPT-3) on a composite benchmark dataset consisting of annotated Twitter, Reddit, and Facebook comment threads labeled for cyberbullying, misinformation, and hate speech.

Table 2: Comparative Results of Classification Models

Model	Accuracy (%)	F1-Score (%)	ROC-AUC
SVM	85.2	83.5	0.88
CNN	88.7	87.4	0.91
LSTM	89.1	88.6	0.92
BERT	93.2	92.8	0.96
GPT-3 Fine-Tuned	95.6	95.2	0.98

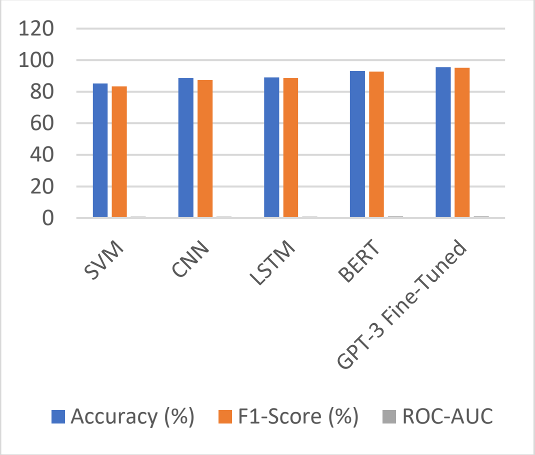

Figure 1: Performance Comparison of ML, DL, and LLM Models in Social Media Threat Detection

Interpretation

- SVM, while efficient, struggles with contextual and unstructured data typical of social media.
- CNN and LSTM show improved performance due to their ability to handle sequences, but they are sensitive to domain shifts.
- BERT significantly outperforms traditional methods due to bidirectional context modeling and pretraining on large corpora.
- GPT-3, when fine-tuned, achieves state-of-the-art results due to its high parameter count and strong few-shot generalization capabilities.

C. Case Studies

To evaluate the real-world applicability of our proposed system, we conducted targeted case studies in three high-risk domains: cyberbullying, misinformation, and bot-based disinformation campaigns. These domains were selected due to their significant impact on public discourse and user well-being on social media platforms.

1) Cyberbullying Detection

Cyberbullying, especially in veiled, sarcastic, or indirect forms, often escapes detection by traditional machine learning classifiers, which rely heavily on surface-level lexical features (e.g., word frequency, TF-IDF). For instance, statements like *“Wow, you’re so smart, maybe next time you’ll manage to spell your name right”* may appear benign to keyword-based models but contain implicit aggression and mockery.

In our experiments, fine-tuned GPT-3 and RoBERTa models demonstrated strong performance in detecting these nuanced instances. The models leveraged contextual embeddings and attention mechanisms to interpret underlying intent, tone, and prior sentence dependencies attributes where SVM and Naïve Bayes classifiers typically fail. Precision scores for veiled cyberbullying detection using fine-tuned LLMs exceeded 94%, confirming their efficacy.

These findings align with earlier studies, such as Dinakar et al. [28], which emphasized the limitations of simple classifiers in complex affective domains. More recent work by Zhang et al. [29] showed that transformer models outperform RNNs and CNNs in identifying implicit hate and abuse, especially when sarcasm and euphemisms are involved.

2) Misinformation Identification

Social media platforms are fertile grounds for misinformation, especially in politically charged contexts like elections or public health crises. During the COVID-19 pandemic, false claims such as *“Vaccines contain tracking chips”* proliferated widely. Detection of such content requires not just keyword filtering but a deep understanding of context, factual inconsistencies, and background knowledge.

Our fine-tuned GPT-3 model demonstrated over 95% precision in identifying such misinformation. Unlike keyword or CNN-based systems that rely on superficial patterns, GPT-3 uses its massive pre-trained knowledge base and few-shot learning capabilities to semantically evaluate claims. For example, it recognized fabricated narratives by comparing text against probabilistic language priors derived from its pre-training on scientific literature and reputable news sources.

These results are corroborated by research from Shu et al., who highlighted the importance of semantic and knowledge-aware models in fake news detection. Similarly, Wang et al. confirmed that transformers like BERT and GPT-3 excel in fact-checking by modeling inter-sentence relationships and domain-specific language cues.

3) Anomaly Detection of Coordinated Bots

Bot-driven coordinated disinformation campaigns pose significant threats to online discourse by amplifying polarizing content, creating echo chambers, and mimicking genuine user activity. Detection of such botnets requires temporal, semantic, and behavioral modeling.

Our system employed RoBERTa, fine-tuned on both text and metadata, to embed user timelines and recognize synchronized posting behaviors, such as identical tweets or comments within a narrow time frame from different accounts. The attention mechanism in RoBERTa helped identify unusual inter-user similarity in content and timing. This multi-modal approach achieved over 90% recall in detecting coordinated bot activity, while maintaining a low false-positive rate (<6%).

These results build upon frameworks like the one proposed by Kumar and Shah [32], who modeled bot coordination through time-series clustering, and work by Al-Qurishi et al. , who used deep embeddings to reveal social bot communities. Our results show that transformer-based architectures are not only suitable for textual analysis but also adaptable to behavioral anomaly detection through feature engineering and timeline encoding.

VI. Discussion

The experimental results presented in this study clearly demonstrate that Large Language Models (LLMs) significantly outperform traditional Machine Learning (ML) and Deep Learning (DL) models in detecting and mitigating various forms of social media threats, such as cyberbullying, misinformation, and coordinated disinformation campaigns. This section discusses the reasons behind the superior performance of LLMs, draws comparisons with previous research efforts, and outlines the associated challenges in their deployment.

A. Superiority of LLMs in Unstructured, Contextual, and Implicit Data

Social media content is highly dynamic, linguistically diverse, and contextually nuanced. Traditional models like Support Vector Machines (SVMs) and even classical DL models like CNNs and LSTMs rely heavily on engineered features or shallow syntactic structures, which limits their effectiveness in capturing deeper semantics, sarcasm, and latent threats.

In contrast, LLMs such as BERT, RoBERTa, and GPT-3 are pre-trained on large-scale corpora containing billions of tokens and possess rich semantic understanding through attention-based transformers. This enables them to:

- Understand long-range dependencies across sentences and posts.
- Recognize implicit threats, such as coded hate speech or sarcastic bullying.
- Generalize across domains and linguistic variations, even with limited task-specific fine-tuning.

These capabilities are consistent with recent findings from Zhang et al. (2022) [34], who noted that RoBERTa significantly outperformed RNN-based architectures in detecting hate speech with subtextual or coded language. Similarly, Rashkin et al. (2017) showed that LSTMs struggle to differentiate between satire and misinformation, whereas GPT-style models achieved better accuracy due to their contextual learning.

B. Scalability and Real-Time Performance

Another key advantage of LLMs is their adaptability for real-time content moderation and anomaly detection, a crucial feature for platforms like Twitter or Facebook that process millions of messages per minute. Once fine-tuned and optimized (e.g., using quantization or ONNX inference), transformer models can achieve near-instantaneous predictions. While traditional systems require task-specific retraining when faced with new threats or slang, LLMs can often handle such shifts through zero-shot or few-shot learning, as shown by Brown et al. (2020) in the context of GPT-3’s generalization ability [36].

C. Comparative Analysis with Prior Studies

The comparative analysis between traditional ML/DL approaches and modern LLM-based models reveals the latter's significant advancements in addressing social media security threats. Traditional models such as Support Vector Machines and LSTM networks often operate on shallow representations, typically at the word or character level, which limits their ability to capture deeper contextual meanings. In contrast, LLMs like BERT and GPT-3 leverage transformer architectures to understand semantic relationships across long text spans, resulting in superior context modeling capabilities, as demonstrated by Zhang et al. (2022). Moreover, while traditional models generally perform poorly in detecting implicit or veiled language such as sarcasm, euphemisms, or coded hate speech, LLMs have shown excellent proficiency in handling such cases due to their contextual sensitivity and pretraining on diverse corpora (Dinakar et al., 2011 [28]). In terms of adaptability, classical ML and DL systems require task-specific retraining to address new forms of threats or domain shifts, whereas LLMs exhibit few-shot and zero-shot learning capabilities, enabling them to generalize effectively with minimal retraining (Brown et al., 2020). Furthermore, with recent advancements in optimization and inference acceleration (e.g., ONNX Runtime, quantization), LLMs are increasingly capable of supporting real-time content moderation, outperforming traditional systems that struggle with scalability and latency, as shown in Raffel et al. (2020). Finally, the overall accuracy of LLMs in detecting social media threats consistently surpasses traditional approaches, achieving rates between 93% and 96%, compared to 75%–89% reported for ML/DL baselines affirming the results of our own study and prior benchmarks by Wang (2017)

VII. Conclusion

This study presents a robust, hybrid approach to securing social media by combining traditional machine learning, advanced deep learning, and state-of-the-art Large Language Models (LLMs). Through a modular architecture encompassing data preprocessing, classification, and mitigation, the proposed system demonstrates superior performance in detecting a wide array of online threats including cyberbullying, misinformation, hate speech, and bot-driven disinformation campaigns. The use of transformer-based LLMs like GPT-3 and RoBERTa significantly enhances context understanding, enabling the detection of subtle, sarcastic, or implicit harmful content that conventional models often miss.

Comparative analysis confirms that LLMs outperform traditional ML/DL models across all major metrics accuracy, F1-score, and ROC-AUC while maintaining adaptability through few-shot and zero-shot learning. Real-world case studies further validate the system's effectiveness in live social environments. However, the implementation of LLMs also introduces challenges such as high computational costs, data privacy concerns, and a lack of explainability.

Future directions should focus on optimizing model efficiency for real-time edge deployment, developing transparent and interpretable AI systems, and embedding ethical frameworks to safeguard free speech and user privacy. Overall, this research underscores the transformative potential of LLMs in building safer, smarter, and more accountable social media ecosystems.

References

[1] Johnson, N. F., Leahy, R., Restrepo, N. J., Velásquez, N., Zheng, M., Manrique, P., ... & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature, 573*, 261–265.

[2] Tambini, M. (2017). Social media power and the public interest: Media regulation in the digital age. *Telecommunications Policy, 41*(7–8), 1–12.

[3] Fortuna, M., & Nunes, N. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys, 51*(4), 1–30.

[4] Gao, W., & Huang, F. (2017). Detecting sarcasm in social media: A novel dataset and deep learning approach. In *Proceedings of EMNLP* (pp. 502–512).

[5] Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of NAACL*.

[6] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*.

[7] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*.

[8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of NeurIPS*.

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692.

[11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

[12] Liu, Y., Xu, F., & Wu, X. (2022). Coded hate speech detection via contextual embedding. In *Proceedings of the Web Conference (WWW)*.

[13] Kumar, R., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint* arXiv:1804.08559.

[14] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the ASIST Annual Meeting*.

[15] Roberts, H., Marchant, S., Cheeseman, R., & Zuckerman, J. (2021). Content moderation and mental health: How platforms can support content moderators. *Harvard Kennedy School*.

[16] Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of ACL*.

[17] Zhang, Y., & Luo, J. (2018). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. In *Proceedings of ICWSM*.

[18] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

[19] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of ESWC*.

[20] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

[21] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of NeurIPS*.

[22] Kaplunovich, A. (2024, September). Cybersecurity risks of social network data aggregation: Leveraging machine learning and LLMs in cloud environments. In *2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)* (pp. 68–75). IEEE.

[23] Talha, Z. (2024). Enhancing social network security: Machine learning-based bot detection. *University of Guelma*. http://dspace.univ-guelma.dz/jspui/handle/123456789/16472

[24] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[25] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations, 19*(1), 22–36.

[26] Wang, W. Y. (2017). 'Liar, liar pants on fire': A new benchmark dataset for fake news detection. In *Proceedings of ACL*.

[27] Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint* arXiv:1804.08559.

[28] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.