AI-Powered Twitter Content Moderation And Alert System: A Comprehensive Survey

Arunthathi R[1], Avanthika D[2], Kailash Nagappan S[3], Surya P[4], Mrs. B. Priyanka[5], Mrs. C. Sangeetha[6]

[1, 2, 3, 4, 5]Dept of Artificial Intelligence and Data Science

[6]Dept of Computer Science and Engineering

[1, 2,, 3, 4, 5, 6]CA Chettinad College of Engineering and Technology, Tamil Nadu, India

Abstract- *Social media platforms such as Twitter (now X) have become primary channels for real-time information exchange, en- abling rapid dissemination of news, opinions, and public discourse. However, this rapid growth has also led to a significant rise in spam, misinformation, abusive language, and malicious content, which negatively impact user experience and platform credibility. Manual moderation methods are no longer effective due to the massive volume and velocity of incoming tweets, making automated content moderation systems essential. This survey examines existing machine learning, deep learning, and transformer-based techniques for de- tecting spam and harmful content on Twitter. It begins with tradi- tional text classification approaches and progresses toward advanced transformer models that provide improved semantic understanding, such as Sentence-BERT (SBERT). Particular emphasis is placed on hybrid deep learning architectures that combine SBERT-based semantic embeddings with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. These hybrid models aim to capture both local textual patterns and long-range contextual dependencies present in tweet streams. The survey also highlights the importance of severity-based content classification, where detected content is categorized according to risk levels to support real- time alerts and administrative monitoring. A comparative analysis of existing approaches reveals performance limitations related to scalability, generalization, and real-time applicability. While hybrid deep learning models demonstrate promising results in multi-class classification tasks, several research challenges remain unresolved. The study concludes by identifying future directions toward developing scalable, accurate, and practical Twitter content moderation systems capable of adapting to evolving online behaviors.*

Keywords: Twitter Moderation, Spam Detection, SBERT, CNN–LSTM, Deep Learning, Alert System

I. INTRODUCTION

The way people exchange information on the internet has evolved significantly. Twitter (now known as X) is one of the fastest-moving platforms for real-time communication in the digital world. Each morning brings waves of brief posts, flowing nonstop as users pass along news, weigh opinions, or simply check in. Though posts shift fast, meaning behind them often lingers longer than expected. A door stands open by choice, inviting honest thoughts and rapid replies. Still, rapid movement incurs expense - uninvited links move fast through these spaces. Right after come fake notifications, damaging statements, deceptive tricks, alongside incorrect information. Even so, hurtful posts can shake a person's confidence and trust in the platform's safety. What shows up on screen sometimes hits close to real-life emotional weight.

Right now, most ways to manage online content rely on humans reviewing posts individually, looking up narrow terms, or sticking to strict templates. Fast-moving floods of information make such slow, human-driven approaches fall behind what's needed now. Things get harder because peo- ple spreading useless messages never stop adapting - they constantly discover loopholes. What happens here doesn't just affect newer systems - it also slows down older ones, especially when work piles up. As things change fast, scientists find themselves leaning hard on combinations like machine learning and natural language processing to handle web text. These setups powered by smart software show up everywhere from filtering comments to sorting feeds.

Spam filters made with simple machine learning relied on familiar cues - words, user history, patterns - alongside standard tools like Naive Bayes or Random Forest. Though better than strict rule-based systems, these approaches often failed when faced with short, relaxed posts, perhaps from Twitter. A change arrived once learning models began finding useful signals in text without being shown how, using layers such as LSTM; they adapted better to nearby or changing data. These days, models such as BERT and Sentence-BERT [1], [2], powered by transformers, deliver stronger insights into Twitter data by shaping words within their context. Since entire sentences carry weight beyond single terms, accuracy tends to improve. Oddly enough, plenty of current setups rely on basic linear blocks tacked onto transformer feeds. That setup feels a bit behind times. Missing patterns might happen when that method is used, especially across sentence boundaries or brief text sections. Certain studies only look at splits between two groups, yet fail to consider what happens with additional categories while assessing threatening content. Peering into Twitter’s response to unsolicited messages shapes this assessment. Not one but two hybrid approaches shine - SBERT paired with CNN, then again with LSTM - each playing its part. Rather than leaning on a single method, views jump between various paths already tested. A few projects zoom in on single issues, making it tough to stretch results elsewhere. It becomes obvious after reading through - better systems must change fast. What’s absent? Alerts that pop without delay during spikes in risky actions.

What’s truly serious doesn’t always get noticed right away. Even though live monitoring helps, its real impact is smaller than you might think. Looking back at earlier results shows missing pieces instead of clear solutions. Right here is where fresh ideas might stretch out - what comes after today’s shifts.

Getting it right still holds weight, though adaptability pushes harder these days.Tools that truly boost web security now open doors once blocked by abuse, false claims, and toxic words. Fast flows of tweets render constant human oversight impossible, pushing dependence on automated systems higher. This work surveys current ways to spot abusive or deceptive posts across Twitter, highlighting shifts in strategy over years. Old approaches leaned on visible word patterns plus hand-coded guidelines- now newer versions prioritize deeper meaning in speech. Focusing shifts toward hybrid setups where Sentence-BERT picks up on word order, then links that to CNNs and LSTMs to map tweet details and timing. Some approaches go further than just labeling data as good or bad, instead ranking severity so teams can act quickly. Looking at actual results, real- world hiccups, and open questions reveals these mixed-method systems tend to strike a steady point - getting things right without losing flexibility. What we see hints at new kinds of moderation tools - ones that adapt quickly, grow easily, yet fit how actual online spaces operate.

A. Core Themes from This Survey

There are four major themes that comprise the key themes of this survey, and these influence how people view things in different ways.

Machine Learning and Deep Learning for Twitter Content Moderation: A closer examination of the process of machine learning, deep learning algorithms, and transformer models used to sort tweets to identify spam, threats, and harm.

Semantic Feature Extraction Using Transformer Models: A closer examination of the process of using Sentence-BERT to extract meaning from the messy data on Twitter, as well as its applications in practical scenarios.

Hybrid Deep Learning Architectures for Improved Classifi- cation: A closer examination of the process of combining deep learning models using SBERT words and CNN and LSTM models to identify regional patterns and step-by-step processes to classify different content into categories.

Severity Analysis and Real-Time Alert Systems: A closer examination of the process of using risk-identification tools such as alert systems and how oversight teams manage them is an effective way to keep harmful content in check.

II. REVIEW OF EXISTING RESEARCH PAPERS

Spam and dangerous posts on Twitter - researchers have looked at many different ways to spot these using machines that learn. At first, people tried old-style text sorting tools built by hand, like Word Clouds, Term Frequency–Inverse Document Frequency (TF-IDF), or short sequences of words repeated. Sometimes extra info got tossed in too - how often someone posted, whether they liked things quickly, or how oth- ers reacted around them. Tools such as basic guesser models, linear separators, or wood-based guessers took turns sorting piles of data. Even though they worked fast and handled basic collections well, these methods struggled to recognize deeper links or context - especially in brief, casual posts online.

As deep learning moves forward, scientists brought neural networks into play to go beyond fixed feature designs. Instead of relying on hand-engineered inputs, they turned to systems like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM), especially useful for handling long strings of text in tweets. Word order matters - models such as Bi- LSTM catch that pattern better than older techniques. Be- cause these networks pay attention to context between words, performance in spotting relevant tweets often rises, marked by higher recall and F1-scores. Still, building these systems needed huge amounts of tagged information. Their responses also shifted easily when noise or uneven data showed up - often seen online.

Twitter spam detection used Convolutional Neural Networks because they can spot small text patterns well. Studies found these networks caught signs like short spam codes, links, or ads accurately. Yet even with strong local detail detection, pure CNNs struggled to understand wider tweet meaning. Long- term context stayed hard for them to handle.

Lately, research has moved toward transformer-driven lan- guage systems including BERT along with Sentence-BERT- these deliver word meanings shaped by context. AraSpam used these meaning vectors, combining them under a multi- purpose deep learning framework - it saw clear gains in sorting text right and performance across new cases. While grasping language meaning well, such methods usually depend on thick network architectures, maybe missing how words flow together or vary locally within sentences. Few modern tools built on transformers handle more than just two-group sorting. Most skip evaluating how severe threats are or triggering alerts altogether.

Looking at what's already been done, deep learning and transformer systems work well for managing Twitter content. Still, there are missing pieces - like models that handle more than two outcomes at once. Data often focuses too much on topics, skips how severe issues are, and rarely includes tools that flag problems quickly. Because of these weaknesses, trying mixtures of semantic meaning with visual-lstm setups makes sense. It could lead to better tools: ones that are strong, flexible, and actually usable for real-world Twitter screening.

A. A Look Back at Earlier Approaches

When you step back and think about it, most of what you find in the literature on spotting bad tweets and spam seems to divide into three categories or buckets—good old machine learning, new stuff like deep learning tricks, and using things like transformers to get the gist of it. How the data behaves seems to change based on what method they use to identify it.

Basically, traditional ML is content to rely on attributes such as TF-IDF, Bag of Words, or n-grams, directly feeding these into models like Naive Bayes, SVM, or Random Forest. They're light on resources and easy to implement – ideal can- didates for first-generation spam filter applications. Although their scope is limited as rule-based systems require human intervention to set rules, which can miss nuanced meanings, especially for casual communications.

In the early days, AI-based systems introduced new models such as CNNs, RNNs, LSTMs, and Bi-LSTMs, all of which learn patterns directly from words and sentences. Moreover, with CNNs, local cues, such as spam words, become easier to recognize. Contrastingly, with LSTMs, word flow over time, such as tweets, becomes easier to recognize. In terms of outperforming the above approaches in terms of precision and detection, these approaches generally outperform them. However, these approaches deeply rely on large quantities of data, and where there is ambiguity or where topics fall out of expected ranges, they can easily go wrong.

This is because, with BERT and Sentence BERT, transformer-based approaches revolutionized the way contex- tual information is processed at a word level as well as a sentence level. This is also due to the fact that these ap- proaches are better at processing meaning compared to CNNs or LSTMs. This shows why transformer-based approaches are currently dominating the leader board for Twitter classification tasks. Yet, much of modern approaches rely on dense layers for text sorting without looking at broader level or local contexts. A large fraction of modern approaches focuses on basic classification, with little attention given to evaluating harm or setting off alarms.

Overall, it seems that these new mixtures are likely more effective, longer lasting, and easier on the fingers when sorting through Twitter.

Table I compares major Twitter content moderation tech- niques based on strengths, limitations, and performance.

III. STRENGTHS AND WEAKNESSES OF EXISTING APPROACHES

1. Strengths

The techniques used today in identifying tweets that are suspected of spam and harmful content include various tools with obvious advantages. Traditional machine learning tools such as Naive Bayes, SVMs, and Random Forests require little resources and are easy to deploy. Such tools are successful with well-structured and well-labeled data and require little time in training; they are thus suitable tools in building a spam detection system.

Deep learning methods, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, are able to learn from raw text data without any need for feature engineering. CNNs are well-suited for detecting local features like spam words, hashtags, and URLs, while LSTMs are able to model the dynamic properties of a stream of words effectively, which in turn improves accuracy.

The transformer-based architectures like BERT and Sen- tence BERT (SBERT) provide sophisticated semantic under- standing using contextual token embeddings. These models perform well in the context of short texts with high noise and informality that are typical of Twitter data, promising excellent generalization across different datasets and topics.

2. Weakness

However, there are certain drawbacks with these approaches. One is that with traditional approaches, there is a heavy dependence on handcrafted features, which do not capture deeper semantic meaning, nor are they flexible with regard to different or broader subjects.

Also, deep learning models require a lot of labeled data and computing power. Although CNN-only architectures fail to link long-range dependencies, LSTMs may be computationally slow and sometimes do not capture certain local features. Moreover, there might be issues related to class imbalance because such a property is predominant in social media data.

Although the usefulness of the transformers is undeniable, there also exist underlying computational costs and deploy- ment difficulties associated with them within a real-time envi- ronment. Many current implementations rely on simple dense classifiers, which do not take full advantage of sequence and regional nuances in text. Additionally, most models still rely on two-class decisions without supporting nuanced severity determination, explainability, and real-time alerting.

These strengths and weaknesses of these systems, individ- ually, highlight the merits of using an effective and hybrid approach, which, apart from utilizing the benefit of semantic learning, also relies on sequence and local feature learning, and facilitates real-time Twitter content moderation.

IV. DEEP LEARNING BASED MODELS

Deep learning-based methods have significantly improved the detection of online spam and harmful content on Twitter. Recognizing messy text patterns becomes easier here because learning happens without needing constant human tuning. Unlike older machine learning styles where setup takes time, these systems adapt on their own far better.

What makes CNNs stand out in sorting texts? They're good at spotting small things - like spammy words, links, tags, or pushy phrases. Instead of sliding filters over plain words, they work on vectors that capture word meanings. These models pay attention to patterns such as sequences of nearby words, called n-grams. Because of this, telling spam from normal messages gets easier when those hidden structures get noticed.

When it comes to handling strings like a tweet, some networks are built for flow - RNN being one. Among them, LSTM stands out by tracking shifts in thought across time. Because it remembers distant signals well, it works better with grammar or subtle phrasing. Looking both ways - backward and forward - Bi-LSTM turns out stronger at guessing next words. Still, handling tough inputs needs strong computing muscle. Learning from vast amounts of labeled examples makes things even tougher.

Even though the deep learning systems work much better than older techniques, there's still space to improve how they show results and connect with other smart tools.

TABLE I Comparison of Twitter Content Moderation Techniques

Technique	**Strengths**	**Limitations**	**Performance Characteristics**	**Applications**
Naive Bayes (NB)	Simple,fast,lowcom- putational cost	Poor semantic under- standing, sensitive to noisy data	Accuracy: 70–78%	Basicspamfiltering
Support Vector Machine (SVM)	Effective with small datasets	Notscalableforlarge datasets	Accuracy: 75–82%	Binarytweetclassi- fication
Random Forest (RF)	Robusttonoise	Requiresmanualfea- ture engineering	Accuracy: 80–85%	Spamdetection
CNN	Excellentlocalfeature extraction	Cannotcapturelong dependencies	Accuracy: 84–88%	Patterndetection
LSTM / Bi-LSTM	Capturessequencecon- text	Hightrainingtime	Accuracy: 85–89%	Contextanalysis
BERT/SBERT	Strongsemanticunder- standing	Computationally expensive	Accuracy: 88–91%	Semanticclassifica- tion
SBERT+CNN-LSTM	Combinessemantic+ sequential learning	Complex architecture	Accuracy: 90–94%	Multi-classmodera- tion

V. HYBRID MODELS COMBINING MULTIPLE TECHNIQUES

Combining multiple machine learning and deep learning methods often leads to improved performance, as they tend to work better together than alone. For example, semantic embeddings can be used to feed information into neural networks inside a system that handles abusive tweets. One idea links with another, creating stronger patterns from raw data.

One well-known mix-of-methods method uses transformer- driven word embedding tools - such as Sentence BERT (SBERT) - to extract meaningful signals, feeding them into classification systems rooted in deep learning, including CNNs and LSTMs. Meaning within tweet sentences often comes through effectively from SBERT's context-aware word layouts. From there, visual scanning tools like CNNs pick up on nearby language clues, while recurrent networks handle longer-term flow. When it comes to spotting fake tweets, using both image and sequence models together often works better than using just one kind of model alone. That's because hybrid systems - like CNN-LSTM setups - can pull insights from words, their layout, and small visual details inside a tweet at once. Studies show these combos tend to hit better marks on accuracy, correct predictions, detection rate, and an average performance metric called F1-score. These gains show up most clearly when sorting through varied labels like spam, genuine posts, or harmful messages. Surprisingly, they also tend to adapt fairly well across different subject areas people discuss online.

As shown in Fig. 1, the proposed system integrates SBERT embeddings...

VI. DISCUSSION ON REAL-WORLD APPLICABILITY

Machine learning and deep learning in content moderation on Twitter are quite pertinent to the real world considering the fact that more and more tweets are filled with spam, misin- formation, and other types of toxic content. Therefore, with the help of automated systems, monitoring of tweets round the clock can be achieved, thus contributing to Twitter's real-time implementation in these ever-changing online environments.

1. Real-Time Twitter Spam and Harmful Content Detection

To be precise, the hybrid method involving SBERT and CNN, LSTM has not only been great but also relevant during instances like live moderation of Twitter since they are capable of grasping the semantic meanings while, at the same time, identifying the context and sequence. They can automatically convert tweets into spam, non-spam, and malicious types with high levels of accuracy.

2. *Severity Assessment and Alert Generation*

There is no doubt that a system response of the same degree is not necessary for all kinds of harmful content. A variety of types of harm using a tier system can be a way of making tweets into low-, medium-, and high-risk categories.

High-risk tweets may prompt system messages to admin- istrators or concerned authorities. This could speed up the reaction and resolution and reduce the manual tasks.

3. *Administrative Monitoring and Decision Support*

Linking moderation models with administrative dashboards could deliver visual insights into spam behavior and prediction probabilities.

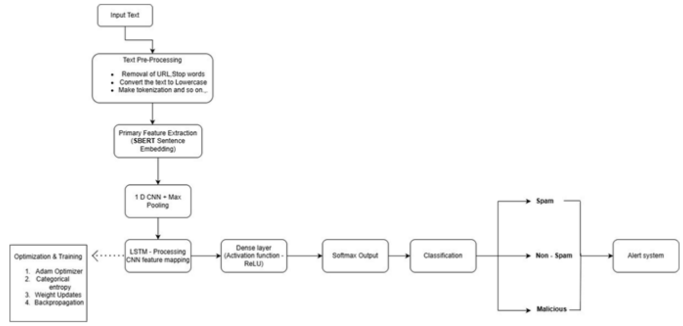

Fig. 1. Proposed hybrid SBERT–CNN–LSTM architecture for Twitter content moderation

VII. CURRENT LIMITATIONS

1. *Real-Time Data Processing Challenges*

Most Twitter spam and harmful content detection systems struggle with real-time processing because of the huge number and speed of tweets. Delays in collecting data, preprocessing, and model inference can lead to late detection, allowing harmful content to spread before any action is taken.

2. *Limited Contextual Understanding*

Traditional machine learning methods that depend on TF- IDF or word-level embeddings fail to capture the meaning and context in tweets. Even some deep learning models do not fully understand sarcasm, implicit hate speech, and the evolving slang often found on social media.

3. *Data Scarcity and Bias*

Many publicly available datasets are small, focused on specific topics, or biased toward certain events like COVID-19. This limits how well models can generalize and makes them less effective when used across various domains, languages, and user behaviors.

4. *Inadequate Multi-Class and Severity Classification*

Several existing systems only classify content as binary (spam vs. non-spam) and do not address different harmful categories, including malicious, abusive, or misinformation content. Furthermore, the lack of severity-level assessment weakens the system's ability to prioritize responses.

5. *Ethical, Privacy, and Transparency Issues*

Automated moderation systems can unintentionally target specific user groups because of biased training data. In ad- dition, the lack of explainability in deep learning models makes it hard to justify decisions. This raises concerns about transparency, user trust, and compliance with data protection laws.

VIII. RESEARCH OPPORTUNITIES

1. *Advanced Semantic Representation Techniques*

Future research can look into better transformer-based embeddings and domain-adaptive SBERT models that can understand informal language, abbreviations, emojis, and mul- tilingual content in tweets.

2. *Multi-Class and Risk-Aware Classification*

Expanding classification frameworks to include multiple categories like spam, malicious, abusive, and misinformation content, along with severity scoring, can enhance moderation effectiveness and response prioritization.

3. *Explainable AI for Content Moderation*

Using Explainable AI (XAI) methods can help visualize attention scores, important words, and decision logic. This will enable administrators and platform moderators to understand and trust automated predictions.

4. *Cross-Platform and Transfer Learning Approaches*

Research can aim to develop models that work across mul- tiple social media platforms using transfer learning, reducing the need for platform-specific datasets.

5. *Human-in-the-Loop Systems*

Combining automated models with human feedback can enhance long-term performance by refining classification ac- curacy and adapting to new threats.

XI. EMERGING TECHNOLOGIES

1. *Edge and Cloud-Based Hybrid Deployment*

Using lightweight models at the edge for initial filtering, together with cloud-based deep models for thorough analysis, can lower latency and improve real-time moderation.

2. *Generative AI for Data Augmentation*

Generative models like GANs and Large Language Models (LLMs) can create synthetic tweets to balance datasets, tackle data scarcity, and strengthen defenses against rare or new spam patterns.

3. *Graph Neural Networks (GNNs)*

GNNs can represent relationships between users, tweets, hashtags, and URLs, helping to identify coordinated spam campaigns and bot networks.

4. *Federated Learning for Privacy Preservation*

Federated learning allows decentralized model training across different platforms or regions without sharing raw user data, improving privacy and compliance with regulations.

5. *Blockchain-Based Secure Logging*

Blockchain technology can ensure the integrity and trace- ability of moderation decisions, alerts, and evidence logs, sup- porting accountability and auditability in content moderation systems.

X. REAL-TIME APPLICATIONS

1. *Real-Time Twitter Monitoring and Dashboard Visualization*

A Flask-based web application can monitor tweet streams, classification results, severity levels, and alert statistics in real time through an interactive admin dashboard. Key challenges include managing API rate limits, addressing data latency, and ensuring system reliability.

2. *Alert Generation and Response Mechanisms*

Automated alert systems can inform administrators or mod- erators when they detect high-risk content. However, minimiz- ing false positives and ensuring alerts are actionable remain significant challenges.

3. *Scalability and Performance Constraints*

Scaling the system to manage large tweet volumes during peak times requires efficient model optimization, load balanc- ing, and resource management.

4. *Policy, Governance, and Compliance*

Successful deployment needs to align with platform poli- cies, legal requirements, and ethical guidelines. Data-sharing agreements, user privacy protections, and moderation trans- parency are crucial for long-term adoption.

5. *Model Maintenance and Adaptability*

Language on social media changes quickly. Continuous retraining, updating datasets, and monitoring performance are vital to maintaining classification accuracy and adapting to new spam and harmful content strategies.

XI. SUMMARY OF KEY FINDINGS

This review focuses on the latest developments in the automation of Twitter content moderation, particularly in de- tecting spam, malicious, and harmful content using machine learning and deep learning. The key points of this review are as follows:

- Embeddings such as those based on transformer archi- tectures, such as Sentence SBERT, provide excellent semantic understanding of tweets, thereby helping in the detection of context-aware spam and malicious tweets.
- Hybrid setups of deep learning systems, especially CNN- LSTM architectures, can effectively capture local word patterns as well as sequential dependencies in sentences, resulting in higher classification accuracy with greater generalization capabilities across a wide variety of tweets.
- Multi-class classification, incorporating different levels of severity, improves the prospects of a more realistic risk- based approach, as opposed to straight-forward binary classification approaches to detecting spam messages.
- Preprocessing stages such as cleaning noises, processing emoticons, and regularizing hashtags enhance perfor- mance by removing extraneous data.
- Web platforms and administration dashboards facilitate real-time monitoring and transparent operation.

In summary, the combination of SBERT with hybrid CNN- LSTM models appears promising for developing accurate, scalable, and deployable Twitter content moderation systems.

XII. FUTURE DIRECTIONS

1. Final reflections on progress within the field

Recently, content moderation on social media has evolved quickly from basic rule-based systems and traditional machine learning technology to highly advanced technology involving deep learning and transformers. It has significantly improved our capacity to detect subtle harms in the context of massive social media systems.

Semantic embeddings like SBERT allow systems to under- stand the content of tweets beyond keywords, while hybrid neural architectures help to deal effectively with the intricacies of language. Real-time dashboards and alerts also assist in making effective use of these models in the day-to-day running of operations.

Yet despite these advances, there are important challenges around bias in data, explicability, privacy, and scale that still need to be tackled. Overcoming these hurdles is a challenge to achieve not only a high degree of accuracy within moderation tools but also a high degree of fairness and transparency for all parties involved.

2. Strengthening real-time detection

Future directions in this work are focused on improving the model inference processes to provide real-time tweet analysis with less latency, hence offering a response to harmful content.

2. *Explainable and Trustworthy AI*

The key will be to build moderation models that can provide a rationale, an understandable explanation, for their decisions.

3. *Multilingual and cross-domain moderation*

Aspects like multiple languages and various cultural settings will make these systems more useful in global social media environments.

5. Adaptive and continual learning

Using continuous learning techniques can also reduce the need for these models to restart every time they need to adapt to changing spam techniques, words, and other abuses.

6. Alignment with platform governance

The new systems should be in line with platform policy and any relevant regulations in terms of ethics, privacy, and moderation.

XIII. CONCLUSION

By combining machine learning, deep learning, and seman- tic language modeling, there exists an opportunity to prevent the proliferation of harmful information shared across social networks like Twitter. This is possible by bringing together technologies like SBERT, CNN-LSTM, and real-time alerts, which can provide accuracy, context, and realism for such moderation techniques. Despite the technical and ethics hur- dles that need to be overcome, ongoing efforts are underway, and technologists, platform teams, and policymakers continue to come together to facilitate the development of better, more responsible, and transparent solutions for content moderation with the prospect of utilizing AI-supported solutions that are far more promising for creating a healthier internet environ- ment.

REFERENCES

[1] Arabic Fake News Detection on X (Twitter) Using Bi-LSTM Algorithm and BERT Embedding, IEEE Research Article, 2025.
[2] “AraSpam: A Multitask Deep Neural Network for Spam Detection,” International Journal of Advanced Computer Science and Applications (IJACSA), 2025.
[3] “TweetGuard: Combining Transformer and Bi-LSTM Architectures for Fake News Detection in Large-Scale Tweets,” International Journal of Data Science and Analysis, Science Publishing Group, 2025.
[4] Real-Time Twitter Spam Detection and Sentiment Analysis Using Ma- chine Learning and Deep Learning Techniques, 2025.
[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on Twitter,” Proc. CEAS, 2010.
[6] K. Lee, B. D. Eoff, and J. Caverlee, “Seven months with the devils: A long-term study of content polluters on Twitter,” Proc. ICWSM, 2011.
[7] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” Proc. ACSAC, 2010.
[8] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary, “Towards online spam filtering in social networks,” Proc. NDSS, 2010.
[9] C. Yang, R. Harkreader, and G. Gu, “Analyzing and detecting spam on Twitter,” Proc. ACSAC, 2011.
[10] A. H. Wang, “Don’t follow me: Spam detection in Twitter,” Proc. SECRYPT, 2010.
[11] N. J. Abdelhamid, A. Ayesh, and F. Thabtah, “Spam detection in Twitter: A machine learning approach,” Proc. ICITST, 2014.
[12] M. Khan, S. H. Khan, and M. Ahmad, “A comparative study of spam detection techniques in social networks,” International Journal of Computer Applications, 2014.
[13] M. Z. Alom, T. M. Taha, C. Yakopcic, et al., “The history began from AlexNet: A comprehensive survey on deep learning approaches,” arXiv preprint arXiv:1803.01164, 2018.
[14] Y. Zhang, Q. Jin, and R. Zhou, “Understanding bag-of-words model: A statistical framework,” International Journal of Machine Learning and Cybernetics, 2019.
[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots,” Proc. WWW, 2017.
[16] A. Gupta, H. Lamba, and P. Kumaraguru, “Hybrid deep learning model for Twitter spam detection,” 2020.
[17] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” Proc. ACL, 2012.
[18] J. Zhang and Y. Luo, “Twitter spam detection using combined features,” IEEE Access, 2019.
[19] A. Alomari, B. ElSherif, and K. Shaalan, “Twitter spam detection: A systematic review,” IEEE Access, 2020.