Review

Systematic Review of Machine Learning and Deep Learning Approaches for Spam Detection in Online Social Networks

Abdu Ibrahim [1,*], Badamasi Imam Ya'u [2,*], Fatima Umar Zambuk [1], Shehu Adamu [3] and Augustine Ndudi Egere [3]

[1] Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi, 740272, Nigeria
[2] Department of Computer Science, Islamic University in Uganda, Mbale, 2555, Uganda
[3] Department of Computer Science, Federal Polytechnic Bali, Taraba, 672101, Nigeria
* Correspondence: abdu01.pg@atbu.edu.ng (A.I.); biyau@iuiu.ac.ug (B.I.Y.)

Abstract

Malicious activities such as phishing, disinformation, and bot attacks for spamming in online social networks (OSNs) have become an increasing problem, and online spam detection is an ongoing challenge to keep users safe and OSNs healthy. However, classical machine learning (ML) algorithms like Random Forest (RF), Decision Trees (DT), k-Nearest Neighbors (KNNs), and Naïve Bayes (NB) have various limitations, including class imbalance, overfitting, and scalability, in the context of OSN environments. Emerging deep learning (DL) algorithms, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, are more tolerant of high dimensionality and temporal data. With recent developments, Graph Neural Networks (GNNs), attention mechanisms, and transformers, along with ensemble learning, are being used to model structural and semantic spam patterns. It presents the results of a systematic review of 25 peer-reviewed papers published between 2020 and 2025 and discusses them in terms of approaches, performance assessment, and limitations. The review found that hybrid DL approaches generally achieved better spam detection performance than traditional machine learning (ML) methods, although at the cost of increased computational complexity and resource requirements. However, since 2020, no systematic synthesis has examined the trade-offs between the performance and efficiency of hybrid models. We observed that hybrid models that integrate GNNs, attention mechanisms, and optimization techniques have great promise for future research and can be used to provide real-time and scalable spam detection for multiple social platforms.

Keywords: Online social networks; spam detection; machine learning; deep learning; graph neural networks; transformer models; explainable artificial intelligence.

Citation: Ibrahim, A., Ya'u, B. I., Zambuk, F. U., Adamu, S., & Egere, A. N. (2026). Systematic Review of Machine Learning and Deep Learning Approaches for Spam Detection in Online Social Networks. Impact in Computics, **2**, 4. https://doi.org/10.65500/computics-2026-004

Received: 21 April 2026| Revised: 29 May 2026 | Accepted: 16 June 2026 | Published: 28 June 2026

1. Introduction

Social networking sites (SNSs) like Facebook, Twitter, and Instagram are significant aspects of contemporary communication, helping to facilitate the quick dissemination of information and interaction, irrespective of distance. But these platforms are increasingly being used for nefarious activities like spam, phishing attacks, fake accounts, and misinformation campaigns. According to recent cybersecurity publications, Phishing, Spam, and Social Engineering attacks spread via social media remain

to have large-scale financial and societal impacts globally [1,2]. Furthermore, social network platforms have been demonstrated to be great tools for spreading fake news or influencing public opinion, as has been the case with the recent global pandemic and geopolitical conflict [3,4]. These advances have necessitated the use of intelligent, scalable spam detection and detection solutions capable of handling the ever-evolving nature of spam in the dynamic Internet "social network" context.

The early spam filters included mostly rule-based filters, and classic machine learning (ML) techniques such as k-Nearest Neighbors (KNNs), Naïve Bayes (NB), Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVMs). Although they successfully uncovered easily detectable spam patterns, they have failed to adapt to the more sophisticated and dynamic spam attacks, as they carry hand-crafted features and were not very successful at capturing complex semantic relationships in textual information[5,6]. As spam became more disturbing and evolved, researchers evolved to more sophisticated learning theories that could automatically establish relevant features from unstructured text in social media.

With massive data volume, the research for spam detection has been revolutionized by a new research paradigm that is Deep Learning (DL), which can learn hierarchical, contextual, and temporal representations. At the same time, the performance of spam classification and contextual understanding has shown significant improvement with the use of models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures [7]. Recent studies have proposed using Graph Neural Networks (GNNs) and Hybrid Ensemble Architectures (HEAs) to exploit the capabilities of combining textual, behavioral, and relational data, which would lead to a more robust and adaptable general system for spam and bot detection in complex Online Social Network scenes [8,9].

This is a systematic review, where we critically discuss the recent articles and research papers on spam detection techniques published from 2020 to 2025 in OSNs. The review collates a wide range of approaches including classical machine learning techniques, DL architectures, graph-based approaches (GNNs) and Graph Convolutional Networks (GCNs), and Graph Attention Networks (GATs), Transformer Based models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), ensemble hybrid models and optimization techniques such as hyperparameter optimization, Reinforcement Learning (RL), Genetic Algorithms (GAs), and Shuffled Frog Leaping Algorithm (SFLA). It also discusses papers that mix specialist DL architectures, such as Feed-Forward Neural Networks (FFNNs) and Generative Adversarial Networks (GANs), with optimization methods to boost spam detection accuracy. Additionally, the study puts into perspective the limitations and merits of these methods, while pointing out current research gaps and opportunities to build scalable, robust, credible, and efficient spam filtering systems for modern online social networks.

Research Questions

The following research questions were used to guide the review:

- RQ1: What is the comparative performance (Accuracy, F1-score, AUC) of traditional ML, DL, and hybrid approaches for detecting OSN spam on benchmark datasets?
- RQ2: What is the performance (accuracy), and training and inference efficiency (time to train and inference) of the latest deep learning models such as GNNs and Transformers?
- RQ3: How well have existing studies tackled the issues of explainability (XAI) and robustness?
- RQ4: What are the key issues with current datasets and how do they affect model generalization?

2. Methodology

2.1 Search Strategy

Following the guidelines developed by [10], a systematic search strategy has been applied to ensure all relevant advances in the field of spam detection in OSNs have been included. It also included journal articles and conference papers published from January 2020 to December 2025, from peer-reviewed literature.

Specifically, the period was chosen to highlight the fast progress of benchmark architectures in Spam Detection, such as DL, Transformer-based methods, GNNs, and Hybrid Ensemble methods. The chosen timeline also aligns with the rise of attention mechanisms, Transformer models, and graph-based learning techniques related to the widespread use of Transformer architectures and GNNs in the field of intelligent detection systems [10], [11], as well as social network analysis and Natural Language Processing (NLP). The review highlighted various studies focusing on the area of spam detection in OSNs, which include ML, DL, GNNs, Transformer-based

models, Hybrid ensemble models, Optimization methods, and Adversarial Learning approaches.

The following key electronic databases were searched:

- IEEE Xplore (2)
- SpringerLink (5)
- ScienceDirect (Elsevier) (9)
- ACM Digital Library (9)
- Scopus (8)
- Google Scholar (10)
- Wiley Online Library (2)

The search strategy employed Boolean operators (AND, OR, and NOT) to combine keywords related to online social networks (OSNs), spam detection, and machine learning/deep learning (ML/DL). The search query was adapted for each database. An example search string is presented below:

("spam detection" OR "spam classification" OR "fake accounts") AND

("social media" OR "social networking sites" OR "Twitter" OR "Facebook") AND

("machine learning" OR "deep learning" OR "graph neural network" OR "attention mechanism" OR "transformer" OR "Bayesian optimization")

This module targets the most relevant keywords associated with the field. In this module, focus is given on some of the most relevant keywords used in relation to the field.

A manual search was also undertaken using the snowball method, where reference lists of studies identified in the database search were checked for further relevant papers that were not located in the database search. It was a combination of these strategies that enabled a thorough search of the available papers.

2.2. Eligibility Criteria

To provide consistency and relevance of the studies, inclusion and exclusion criteria were set.

2.2.1 Inclusion Criteria

- Journals and conference publications (2020 - 2025).
- Research focusing on OSNs (such as Twitter, Facebook, Instagram, LinkedIn, and YouTube) that explicitly dealt with spam.
- Applications/Test of ML, DL, or hybrids for Spam/Malicious Spam (Bot accounts, phishing, etc. "spam-like" activity).
- Papers that have measurable outcomes (accuracy / precision / recall / F1 score / AUC, etc.).
- Papers in English

2.2.2 Exclusion Criteria

- Pre-2020 and post December 2025 publications.
- Grey Literature, Preprints, Technical Reports, Theses and Dissertations.
- Studies only on rule-based approaches not including ML or DL.
- Research activities which are not directly related to OSNs
- Review papers, surveys or purely theoretical studies.
- These criteria have provided a filter for inclusion of only empirical and rigorous research in this synthesis.

2.3 Review Protocol and Study Selection Procedure

To improve methodological transparency and reproducibility, the review process followed the PRISMA 2020 guidelines throughout the study selection and screening stages. Although this review was not prospectively registered in a formal review registry such as PROSPERO because it focuses on computer science and artificial intelligence literature rather than clinical interventions, a predefined review protocol was developed before data collection. The protocol specified the research questions, search strategy, eligibility criteria, quality assessment criteria, and data extraction procedures.

The screening process was conducted independently by two reviewers (the lead researcher and supervising author). Titles and abstracts retrieved from the selected databases were first screened against the predefined inclusion and exclusion criteria. Subsequently, full-text articles were assessed for eligibility. Any disagreement regarding study inclusion was resolved through discussion and consensus. Where consensus could not be reached initially, the article was re-evaluated jointly using the predefined criteria until agreement was achieved.

To enhance reproducibility, all screening decisions, quality assessment outcomes, and extracted study characteristics were documented systematically throughout the review process. The complete screening and selection workflow is illustrated in the PRISMA flow diagram.

2.3.1 Study Selection Process

In all the database searches, 642 articles were found. After duplicate articles were removed (using EndNote reference management software), 498 articles remained. The three stages in the screening process were:

- Based on the title and abstract, we eliminated articles that were obviously not relevant (e.g., medical misinformation, ecommerce recommendation systems), with 168 articles remaining.

- Full-Text Screening: The remaining articles were screened for eligibility. For example, 96 articles were rejected because they lacked methodology, performance measures, and/or used non-OSN data.
- Snowballing - citations of eligible articles were searched for additional relevant studies (12 studies added).

After the screening process, 84 studies satisfied the inclusion criteria. To ensure a rigorous and focused synthesis, these studies underwent an additional quality assessment based on methodological quality, dataset relevance, evaluation robustness, innovation, explainability, and scalability considerations. Each study was evaluated using predefined quality criteria and assigned a quality score. Research with higher methodological rigor, applicability to online social network spam detection, and adequate data diversity and approach diversity were prioritized for inclusion. Therefore, 25 studies were chosen for the final synthesis. The detailed quality assessment procedure and scoring results are presented in Appendix C. A flowchart of the screening process (PRISMA) is provided in Figure 1.

Figure 1 illustrates the PRISMA 2020-based study selection process. Records were identified through systematic database searching across seven electronic databases and supplemented by citation snowballing of reference lists of eligible studies. Following duplicate removal, records were screened through title and abstract evaluation. Eligible full-text articles were further assessed using the predefined six-criteria quality assessment framework before final inclusion in the synthesis.

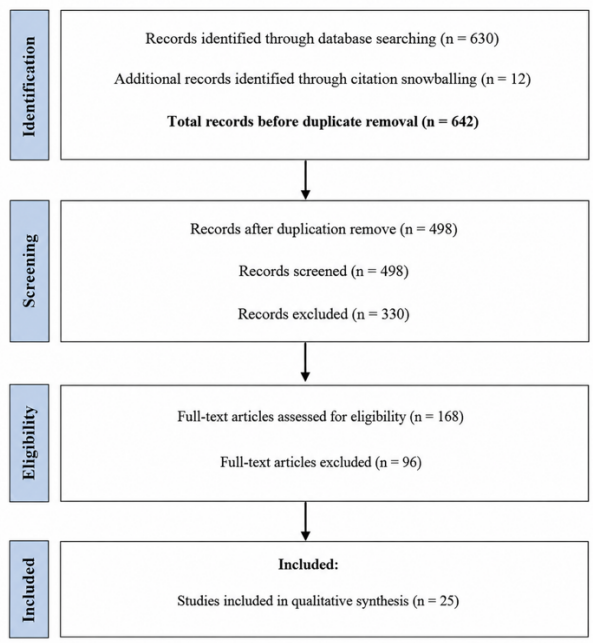

Figure 1. PRISMA flow diagram of study selection process.

2.4 Data Extraction

All the studies were utilized by applying in advance a predefined data extraction form. To respond to the research questions, the following information was extracted (Table 1).

Table 1. Research question Metric extraction

Field	Description	RQ Addressed
Bibliographic Details	Author(s), year, source of publication	–
Proposed Methods	ML, DL, hybrid, or ensemble techniques used	RQ1
Datasets	Type, size, source, class balance, public availability	RQ4
Metrics	Accuracy, precision, recall, F1-score, AUC	RQ1
Computational Resources	Time to train, hardware, model size	RQ2
Explainability Approaches	SHAP [13], LIME [14], Attention, etc.	RQ3
Adversarial Robustness	If presented, methods and results for this task	RQ3
Strengths & Limitations	Author's discussion or our assessment	–
Improvement	New (such as, use of GNNs and Bayesian hyperparameter search)	–

This has contributed to extracting all the information in the same way and allowed for comparison of studies.

2.5 Quality Assessment

In line with guidelines for performing systematic reviews in computer science research [10], a quality assessment procedure was carried out to evaluate the methodological qualities and reliability of the reviewed studies. The six dimensions assessed for each study were given scores of 0, 1, or 2 (Not Satisfied, Partially Satisfied, Fully Satisfied) and ranged from 0 to 12 points per study. The six assessment criteria and scoring descriptions are listed in Table 2.

Table 2. Quality assessment framework and scoring criteria used for evaluating the included studies on online social network (OSN) spam detection.

Dimension	Description	0 - Not Satisfied	1 -Partially Satisfied	2 – Fully Satisfied
Data Suitability (DS)	Whether the dataset is diverse, sufficiently large, and representative of OSN spam patterns	Dataset is small, single-platform, or unrepresentative	Dataset is moderately sized or partially representative	Dataset is diverse and well suited to OSN spam detection
Methodological Transparency (MT)	Transparency of research design, model description, and replicability	Methods are poorly described or irreproducible	Methods are partially described with some gaps	Methods are clearly stated and replicable
Evaluation Robustness (ER)	Use of standard evaluation metrics with appropriate validation strategies	Only one metric reported; no cross-validation	Multiple metrics reported but validation is limited	Comprehensive metrics (Accuracy, F1, AUC) with robust validation
Innovation & Contribution (IC)	Added value to the existing knowledge of OSN spam detection	Marginal or no novelty over prior work	Moderate novelty with incremental improvements	Clear and significant methodological or applied contribution
Explainability (EX)	Extent to which the study addresses model interpretability	No interpretability addressed	Partial interpretability (e.g., attention visualization)	Explicit XAI methods employed (e.g., SHAP, LIME)
Scalability Consideration (SC)	The study discusses real world deployment or computational efficiency	Scalability not discussed	Scalability partially discussed	Scalability explicitly evaluated or addressed

Results scored 10-12 were considered High Quality, 7-9 Moderate Quality, and 0-6 Low Quality. Studies with ratings of Moderate or High Quality only were considered for the final synthesis.

Of the 84 studies that were evaluated, 59 were excluded due to not having met the quality criteria outlined in the beginning. More detailed quality scores for all the included studies are available in Appendix C. Typical exclusion criteria included absence of robust evaluation measures (such as using only one evaluation measure without cross-validation) and absence of adequate methodology (only one evaluated resource, or no selection process performed).

2.6 Data Synthesis

Two methods of synthesizing the studies were employed: the narrative method and the comparative method. Studies compared a variety of performance metrics such as Accuracy, Precision, Recall, F1-score, AUC, and computational efficiency. The studies were classified according to the synthesis approach as follows:

- Traditional ML methods (RF, DT, KNN, SVM, LR, XGBoost
- Deep learning approaches (CNN, RNN, LSTM, Bi-LSTM)
- Graph-based (GNN, GCN, GAT)
- Transformer-based models, (BERT, GPT, Attention Mechanisms)
- Hybrid and ensemble models
- Optimization and robustness techniques (SOA, RL, GA, SFLA)

Major trends, strengths, limitations, and research gaps in the study of spam detection were determined through narrative synthesis. Although traditional ML models have been interpretable and efficient, they are severely limited by changing spam patterns and high-dimensional data. Large datasets and high demand for computational resources were needed for these DL and transformer-based models to obtain better contextual understanding and accuracy. Graph-based and hybrid models showed enhanced robustness but were not scalable and computationally efficient, especially in real-time OSN environments. This review study did not include any human participants or primary data collection since all reviewed studies were from publicly available peer-reviewed literature.

3. Systematic Synthesis of Literature

The application of ML and DL to spam detection, fake-account identification, and malicious content analysis in online social networks has been widely discussed in recent literature [15]. The following are some papers that have tried to address some of the challenges, such as class imbalance, effectiveness, efficiency, and adapting to new spam types, etc. The findings of this body of research will shed light on innovative techniques as well as challenges still awaiting in this area. The studies have been organized by methodological themes, and comparative discussions

are included within the themes for clarity and a more critical analysis.

3.1. Traditional Machine Learning

In OSNs, the responses of traditional ML classifiers, commonly used as baseline techniques for Spam Detection, are typically known. They are known to be simple, interpretable, and efficient in computation in dealing with small and medium datasets. Their effectiveness, however, is often limited in high-dimensional and highly dynamic OSN environments

Traditional machine learning algorithms such as Random Forest (RF), Decision Tree (DT), k-nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB) were studied by [16] experimentally on the task of detecting spam comments in YouTube. According to their results, traditional ML methods, like RF, were effective in the task of spam detection in social media with almost perfect accuracy and AUC-ROC results. Effective preprocessing and feature engineering are key factors in getting the model to perform well.

The study by [17] proposed a Twitter suspicious content detection framework operating based on traditional ML classifiers and combination feature extraction strategies. Over 90% performance was obtained in the spam detection tasks using experimental studies by ensemble and traditional ML models. As new spam trends appeared, the robustness and generalization of these models decreased.

Similarly, a lightweight spammer detection framework for Twitter using RF, Logistic Regression (LR), and NB classifiers was presented by [18]. The results indicated that improved performance for spam tweet classification can be achieved using ensemble learning and multiclass detection. But as the spam landscape has changed and the new tweet variety has become more diverse, the real-time detection accuracy remained unstable.

Traditional ML classifiers based suspicious content detection framework for Twitter was developed by [17], combining various feature extraction techniques such as CNNs. It was observed through experiments that traditional ML and ensemble models obtained more than 90% performance in spam detection tasks. The models, however, were impacted by model robustness and generalization, due to their dependency on the quality of the features and the evolution of spam techniques.

Overall, all these traditional ML models are an important benchmarking standard in the research of spam detection. However, they cannot represent complex relations, contexts, and temporality that can be found in contemporary OSNs. They are less effective when used in modern large-scale OSN spam detection systems due to their limited scalability and adaptability towards advanced multi-channel adversarial spamming technologies.

3.2. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) Networks

Recent deep learning-based methods are more successful in learning hierarchical and sequential patterns from raw OSN data, which significantly reduce the manual feature engineering efforts.

In their research, [19] explore deep learning techniques for detecting posts that are spam on the social platforms, Twitter, for instance, employing CNN-LSTM architectures. The study showed that the semantic and context models extracted from hybrid deep learning methods were efficiently represented and extracted the semantic and context patterns from text data in social media with good accuracy in building the spam detection model even in real time. But the Performance requires large, labelled datasets and high computational costs.

The research CNN-based model for Twitter spam Detection using Tweet-text classification techniques presented by [20], demonstrated that a CNN-based network was very useful to include semantic tweet features and perform the spam-detection task better than other machine-learning techniques. The framework, however, needed a substantial amount of learning time, huge databases, and high computing power.

A hybrid CNN-LSTM-based approach for Twitter spam detection, using semantic representation from WordNet and ConceptNet, was suggested by [20]; their experimental results demonstrated that the accuracy of spam classification and feature extraction was higher than the standalone CNN and traditional ML models. But the hybrid structure needed huge amounts of data, pre-processing, and computational power to be able to deploy it effectively.

Likewise, [22] suggested a social network spam detection system that utilizes A Lite Bidirectional Encoder Representations from Transformers (ALBERT) embeddings and a self-attention Bidirectional Long Short-Term Memory (Bi-LSTM) model to improve contextual spam detection. They tried to show improved accuracy, precision, and F1-score performance in capturing the semantic and contextual relationships in spam text. The

new mechanism of self-attention resulted in higher computations and resource usage, though.

To overcome the issue of class imbalance in the datasets of spam detection on the Twitter platform, the SMOTE-ENN resampling technique with deep learning classifiers was applied by [23], and their results revealed that their approach to balance the number of spam and non-spam samples resulted in higher model performance and stability with respect to various metrics. The researchers also highlighted that synthetic oversampling could diminish generalization because synthetic samples may not reflect the evolving and dynamic nature of real-world spam behavior in the case of OSNs.

Based on these properties, critical analysis: DL models have been proven to have excellent generalization powers and to perform best on complex spam patterns in OSNs. But such models are still computationally intensive and require a large and good-quality dataset. Although attention mechanisms can improve interpretability of models and understanding of context, they are also complex, and complex networks with large amounts of data may not be able to scale up real-time applications.

3.3. Graph-Based Approaches: Graph Neural Networks (GNNs), Graph Convolutional Networks (GCNs), and Graph Attention Networks (GATs)

GNNs have recently been proven to be very effective in spam detection, particularly for OSNs with explicit modeling of the structure and relationships of users' interactions. As a form of graph-based processing, GNNs are particularly well suited to finding organizations or clusters of spammers with similar content that are covertly connected and coordinated attacks by them that would be hard to measure with traditional machine learning.

The study Attention-based Spam Review Classification in Heterogeneous Social Networks framework based on GNN. Built an effective Spam Graph Neural Network (SGNN) that can effectively detect complex relational and semantic spam patterns with good detection performance for both Amazon and YelpChi datasets. For large-scale datasets, though, heterogeneous graph processing and hierarchical attention mechanisms caused models to become more complex and resource-heavy [24].

A fraud and spam detection system in online social networks based on the GCN has been proposed by [25]. The model performed well when it came to detecting anomalies, and it was able to capture relational interaction patterns between users well. The study articulated that the application would have higher computational complexity and lower scalability when working with large amounts of graph data.

Moreover, it proved that it is possible to add attention mechanisms to a GNN for spam bot detection in OSNs [26]. They designed their GAT-based model to better aggregate features and capture important information for spam detection in a noisy social network environment. The study also mentioned problems that occur with noisy and complex graph structures in graph-based spam detection systems.

Critical Analysis: GNNs are well suited to the idea of structural spam detection as they use relational/ network-based information that traditional ML/ DL models do not capture. But there is potential for scalability issues, computational difficulty, and overfitting. Noise and incomplete graph structures exist in real-world OSNs, impeding model robustness and generalization ability greatly.

3.4. Attention-Driven Transformer-Based Models

Although the transformer-based models with self-attention mechanisms are effective for long-range dependence and semantic relationship extraction in spam text, they bear the disadvantage of high computational cost. Transformer-based models with self-attention mechanisms can effectively encode long-range dependencies and semantic relationships in spam text, but they suffer from the problem of high computational cost. Using a network of contextual information and complex linguistic patterns, which leads to more accurate information understanding, they can outperform traditional machine learning and previous deep learning techniques in detecting spam in text-based emails when dealing with OSNs.

To boost the accuracy of spam detection on Twitter datasets, [27] introduced the Attention-based Social Spam Detection Framework (SDF) that jointly utilized BERTweet representation, attention mechanisms, and GAT. They overcame the problem of spam detection with their model by recognizing the importance of tweets and incorporating user interaction information. But the addition of the attention and graph-based learning features added computational complexity and training expenses, making them less scalable for big datasets.

As in the case of [27], which proposed a transformer-based social bot detection framework that combines pre-trained language models (BERT and GPT variants) with a Feed-Forward Neural Network (FFNN) for tweet-level classification. The model achieved an F1-score of 93% and improved accuracy by 3-24% over baseline methods. Fine-

tuning of the transformer models, however, needed electric power consumption and a lot of annotated data.

A study on malicious social spam detection, which incorporated a topic-aware neural attention mechanism with LSTM [28], was conducted, where the researchers demonstrated that their model reached remarkable levels of contextual understanding and robustness in the classification of malicious user spam over traditional deep learning. However, combining multiple attention and semantic models came at a cost of high computation, cost and training overheads.

Critical Reflection: Modelling long-range dependencies and understanding context is crucial in text spam detection using transformer-based models. They, however, suffer from high computational and memory costs and are thus not suitable for resource-limited settings. The current studies will, therefore, focus on lightweight transformers, model compression, knowledge distillation, and efficient attention mechanisms to improve scalability while retaining high detection accuracy.

3.5. Hybrid and Ensemble Models

Hybrid and ensemble models involve the use of several learning architectures that can leverage the merits of each model and generally achieve better spam detection than individual ML or DL models. These models fuse the textual, behavioral, contextual, and structural information to enhance the robustness and accuracy of spam classification in OSNs.

A hybrid spam detection framework for spam review classification, which combines spammer behavioral features and linguistic content features, was suggested by [30]. By the SD-FSL-CLSTM, they achieved remarkable gains in spam detection accuracy by incorporating the interactions between behavioral features and textual features. The challenges that came out of the study included feature selection, processing high-dimensional data, and changing spammer behavior.

A hybrid spam detection architecture for social networks that integrated spam detection machine learning, text analysis, and short link analysis techniques was developed by [31]. Their design integrates multiple spam detection methods, achieving a precision and F-measure of 95.69%. However, there are scalability and processing-time problems when it comes to analyzing large amounts of data in "real-time" environments.

Self-attention mechanisms and deep learning have been proposed to detect social spam on imbalanced data by [8]. They articulated a framework whose introduction received two points: One is for improvement in the performance of spam classification using a multi-architecture deep learning approach, and another is for introducing contextual feature learning techniques for the improvement of spam classification approaches for robustness. However, the hybrid architecture had some expense in computation complexity and some overhead while training and gave some limitations while scaling in real-time performance.

Similarly, the study of a spam detection method based on a hybrid model (Gated Recurrent Unit (GRU), Autoencoder (AE) & Multilayer Perceptron (MLP)) provided by [32] also followed the same approach. Their ensemble approach yielded top performance on various spam datasets, with the use of complementary deep learning techniques. However, incorporating multiple learning components substantially increased computational and training requirements.

A dynamic deep ensemble method that integrates two classifiers, namely the CNN and the RF with the Extremely Randomized Tree classifiers, to identify spam was developed by [6]. They achieved better spam classification performance using a combination of automatic feature extraction and ensemble learning strategies. But the implementation of using multiple deep learning and ensemble parts made the computation complex and resource-consuming.

Thus, recent multimodal hybrid architectures are known as the cutting-edge technique in spam detection research due to their complementary abilities that combine various spam detection models and data modalities. Nevertheless, though they show higher performance, these architectures are still challenging to implement due to high computational complexity, limited scalability, and higher implementation cost.

3.6. Optimization and Robustness Techniques

Optimization and robustness techniques have gained more and more traction to enhance the (anti)spam ability of spam detection models in OSNs. The idea is to increase the generalization of the model, reduce overfitting, and increase adversarial spam tolerance.

For enhanced spam detection in social networks, [33] presented a Reinforcement Learning (RL) based stratified hyperparameter optimization framework called SON2S, which is based on Feed-Forward Neural Network (FFNN) parameter optimization using a modified version of the Shuffled Frog Leaping Algorithm (SFLA). The method led to higher detection accuracy and lower false positive rates than previous methods but also introduced more complexity into computation when optimizing.

Similarly, an optimization approach for spam detection based on identifying malicious URLs using feature selection methods like the Firefly Algorithm and Harris Hawks Optimization coupled with the KNN algorithm was suggested by [34]. Their findings, optimized feature selection improved the accuracy of spam detection in accordance with their proposed model. May have higher computational requirements and lower scalability needs for large data in real-time applications.

The study proposed a hybrid deep transformer model and hyperparameter optimization framework for social media fraud detection in the dataset of Twitter, using Temporal Convolutional Networks (TCN), GAN-based augmentation, autoencoders, and hyperparameter optimization using Seagull Optimization Algorithm (SOA). Demonstrated high robustness, scalability, and high ROC-AUC. However, the framework was quite memory-intensive, had many computational requirements, and complex optimization methods as well [35].

To address feature dimensionality and class imbalance in Twitter spam detection, [36] proposed a modified Genetic Algorithm (GA) using XGBoost for simultaneous feature selection and hyperparameter optimization. The framework was successful in achieving an accuracy of 92.67% but required an extra computational overhead due to the genetic search and repeated cross-validation.

Likewise, the use of GANs to optimize the spam detection problem and overcome the problem of class imbalance when dealing with the social media spam problem was proposed by [37]. Their framework has improved the accuracy of spam classification tasks: CGANS (Spam Generating synthetic Spam Nonspam), which adds synthetic spam examples to the training set. But GAN learning posed the problems of training stability, convergence, and computational complexity.

Hence, robustness and optimization techniques are important and significant for enhancing the generalization ability and robustness of spam detection models against the attacks of adversaries. Yet, the approaches often achieve ease of computation and resource consumption at the expense of robustness and scalability, providing an important performance trade-off between the two.

For ease of comparison, we have created a summary of the 25 reviewed studies, organized to address our research questions, in Table 3. Shows trends, for example, the balance between performance and efficiency, data characteristics, explainability, and robustness.

Table 3. Overview of existing approaches for OSN spam detection and their reported performance and limitations.

Study	Methods Proposed	Results and Conclusion	Metrics Result	Drawbacks
[25]	GCN framework for fraud/spam detection in social networks.	Improved anomaly detection performance.	Accuracy=94.5%, F1-score =93.8%	High computational complexity and scalability issues for large graph datasets.
[26]	Graph Attention Network (GAT)-based spam bot detection framework	Improved feature aggregation and spam detection performance in OSNs.	AUC = 95.4%, Recall = 0.88, Precision = 0.93 and F1-score = 0.91	Noisy and incomplete graph structures reduced robustness and accuracy.
[19]	CNN-LSTM deep learning framework for Twitter spam detection.	Improved contextual spam detection performance in real-time environments.	Accuracy 98% and F1-scores = 0.98.	DL models require large datasets and high computational resources.
[22]	Bi-LSTM integrated with self-attention mechanisms	Improved contextual spam recognition and semantic understanding in spam messages.	Accuracy = 98.42%, F1-score = 98.31%	Higher computational resources and longer training time.
[28]	BERT/GPT-based pre-trained language models with a Feedforward Neural Network (FNN) for tweet-level social bot detection; supported by XAI	Achieved superior bot detection performance by leveraging contextual embeddings from transformer models, outperforming traditional embedding methods	F1-score: 93%**;** Accuracy improvement: 3–24% over baseline methods.	High computational and memory for transformer fine-tuning; dependence on large datasets; longer training time
[18]	Twitter spam detection using NB, LR, and RF with lightweight detectors	Improved spam tweet classification performance	NB: Accuracy = 79% LR: Accuracy 78%, RF: Accuracy =75%	Detection accuracy affected by evolving spam tactics and tweet diversity
[17]	Twitter suspicious content detection using DT, RF, KNN, LR, and NB with hybrid feature extraction	ML and ensemble models effectively detected spam and suspicious Twitter content	Over 90% accuracy, precision, recall, and F1-score achieved	Performance depends feature extraction and struggles with evolving spam behaviors
[27]	Combined BERTweet, attention mechanisms, and (GATs) for Twitter spam detection.	Improved contextual spam recognition and feature relevance in Twitter datasets.	Accuracy: 95.2%, F1-score: 94.1%	Increased computational complexity and poor scalability for large datasets.

[6]	Hybrid CNNs with Random Forest and Extremely Randomized Trees for spam detection.	Improved spam classification through automatic feature extraction and ensemble learning.	Accuracy = 98.38%	Increased complexity and resource requirements.
[33]	SON2S framework combining Reinforcement Learning and modified Shuffled Frog Leaping Algorithm for FFNN hyperparameter optimization	Improved spam detection accuracy and reduced false positive rates on Twitter and Tip Spam datasets compared with existing techniques	Higher Accuracy of 97.5% and Lower False Positive Rate than baseline methods	Computational expensive and optimization complexity due to multi-stage hyperparameter tuning.
[20]	CNN model for Twitter spam detection using tweet-text feature extraction and classification	Improved spam tweet classification by capturing semantic and contextual tweet	Accuracy = 88%, Precision = 93%, Recall = 84% and F1-Score = 88%.	Required large labeled datasets, extensive preprocessing, and high computational training
[23]	SMOTE-ENN with deep learning classifiers	Improved model stability and spam classification on imbalanced datasets.	Accuracy: 99.26%, recall: 99.07 and precision: 99.49	Synthetic oversampling reduced generalization capability.
[31]	Hybrid ML, text analysis, and short-link analysis framework	Enhanced spam account detection performance in social networks.	F-measure and precision = 95.69 %.	Computational complexity and scalability challenges in real-time environments.
[8]	Hybrid ensemble framework combining deep neural networks with self-attention mechanisms.	Improved robustness and spam detection performance on imbalanced datasets.	Accuracy = 97.8%, F1-score = 97.2%	High computational overhead and training complexity.
[38]	Twitter spam detection using RF, K-NN, NB, and DT classifiers	RF achieved the best spam detection performance	RF Accuracy = 92%	Struggle with evolving spam patterns and dynamic tweet behaviors
[24]	Attention-based Graph Neural Network for spam review classification	Captured complex relational spam patterns effectively in heterogeneous social networks.	Accuracy=78.60%, F1-score=79.38%, Precision =0.87	Computationally expensive on large-scale graphs.
[32]	EGMA ensemble framework combining GRU, Autoencoder, and MLP	Achieved strong spam detection performance across datasets.	Accuracy = 99.28%	Longer training time and high resource consumption.
[36]	Modified Genetic Algorithm with XGBoost for feature selection and hyperparameter optimization.	Improved Twitter spam detection by addressing feature dimensionality and class imbalance	Accuracy = 92.67%, G-Mean = 82.32%, using <10% of original features	Additional computational overhead due to genetic optimization and repeated cross-validation
[37]	Applied GAN-based learning for social media spam detection and class imbalance handling.	Improved spam classification performance and robustness.	Accuracy = 98.2% and F1-score = 93.0%	Training instability, convergence issues, and high computational complexity.
[16]	YouTube spam detection using RF, DT, KNN, SVM, LR, and NB classifiers	Random Forest outperformed other traditional ML models	RF: Accuracy = 100% Precision= 0.9841 AUC-ROC score	Performance depends on feature engineering
[21]	Hybrid CNN-LSTM model with WordNet and ConceptNet for Twitter spam detection	Improved feature extraction and spam-classification over standalone CNN and classical ML models	Precision = 94.91%, Recall = 96.76% and F-score = 95.83%	High computational cost, extensive preprocessing, and large datasets
[35]	Hybrid TCN-GAN-SOA deep-learning framework with autoencoder and hyperparameter optimization	Improved fraud-account detection robustness and scalability on Twitter datasets	ROC-AUC = 0.96	High memory usage, computational complexity, and expensive optimization/training
[29]	Neural attention framework integrating LSTM, topic modeling, and BERT for social spam detection	Improved contextual awareness and robustness against malicious social spam	Accuracy = 99.02%, Precision = 98.90%, Recall = 98.87%, F1-score = 98.88%	Increased computational complexity, higher processing cost, and expensive multi-attention training
[34]	KNN with metaheuristic feature selection for spam detection	Improved malicious URL and spam detection performance.	High classification performance Accuracy = 99.76%	Increased computational complexity with larger datasets.
[30]	Hybrid behavioral and linguistic feature fusion	Improved spam classification performance over single-feature methods	Accuracy = 98.7%, F1-score = 98.5%	Increased complexity in high-dimensional environments.

Figure 2 presents the taxonomy of ML and DL approaches identified in the reviewed studies. The taxonomy classifies the approaches into traditional ML, DL, graph-based learning, transformer-based approaches,

hybrid models, and optimization techniques. This classification attempts to capture the progression of spam detection techniques, which have evolved from traditional classifiers to sophisticated graph neural networks, transformer models, and optimized hybrid spam detection techniques to boost spam detection in OSNs.

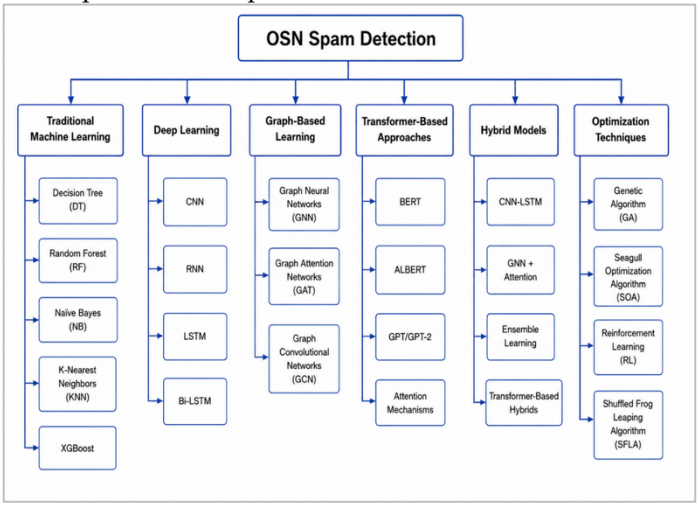

Figure 2. Taxonomy of OSN Spam Detection Approaches

4. Discussion

Considering the years 2020-2025, this systematic review critically captures the latest research trends in spam detection methods in OSNs along with different categories of techniques including ML, DL, GNN, Transformer-based, hybrid ensemble, and optimization-driven approaches. The results witness clear migration from classical feature-based ML algorithms to more recent context and graph-based DL models and algorithms to solve the dynamic and adversarial spam activities in the present social platforms.

Based on the results of the review, we can say that the traditional machine learning models like RF, DT, LR, Naive BN, and KNN work with less amount of computational power and are also simple, particularly for a smaller and well-balanced dataset. The accuracy of the spam detection was corroborated by [16], [37], and [17], who demonstrated that the accuracy of the spam detection was close to, and in some cases even exceeded, 90% using the RF-based methods. Each of these, however, had drawbacks since they were primarily composed of hand-designed features and superficial textual representations, all of which are not able to follow the development of spam strategies, semantic manipulation, and the counterattack on content generation. It is similar to RQ1, introduced to compare and gain understanding of the efficiency of both ML and DL techniques in detecting OSN spam.

In recent times, due to the transformation of the Spam Detection System to deep learning, there has been an improvement in the understanding of context and automatic feature extraction of the system. Training a CNN, RNN, LSTM, and Bi-LSTM was found to be more effective in learning the semantic and temporal relationship of the data in OSN. Classes achieved by them are very high, with some leading to an F1 score of over 95% [18], [20], and [22], and some reported by [20] to be above 95%. The following are architectures that helped to minimize manual feature engineering and increased their versatility with the ever-changing spam pattern. The DL approaches, however, needed a significant amount of pre-processing, a considerable amount of labeled data, as well as a lot of computing power to provide consistent results. Hence, such improvement in the classification capability would be achieved with the DL method, and this would also be a scalable and real-time OSN deployment problem.

The review also highlights and focuses on the research efforts undertaken by both GNNs and graph attention approaches. The difference between traditional classifiers based on text and the potential power of GNN's is that they can include the structural information from social interactions, such as who accepted or rejected an incoming communication, identification of coordinated spams, Bot communities, and malicious network behaviors. The studies by [25,26]and [24] demonstrated the effectiveness of graph-based models in enhancing the accuracy of anomaly detection and spam classification by representing user interactions as a graph structure. But graph processing added significant computational overhead, scalability limitations, and noisy or incomplete graph data. The results directly answer RQ2 about the compromise between detection performance and computational efficiency of advanced deep learning architectures.

Transformer-based models and attention models are some of the most effective models used to understand contextual spam in OSNs. The inclusion of BERT, GPT, BERTweet, self-attention, and topic-aware attention mechanisms improved the ability of the models to represent long-range dependencies and semantic relationships better than previous DL architectures. The attention-based approaches significantly enhanced the robustness of contextual spam detection, and the semantic interpretation based on [27,28] and [29]. For instance, [29] used LSTM, BERT, and topic-aware attention mechanisms to achieve an accuracy of 99.02% and an F1-score of 98.88%. These models, however, needed to consume huge amounts of memory, involve cumbersome fine-tuning methods and

demanded huge computational power, which made them impractical in resource-constrained or real-time OSN applications.

One of the other key trends that emerged from this review is the increased use of hybrid and ensemble architectures. A hybrid framework integrates several learning paradigms to take advantage of the complementarities among the textual, contextual, behavioral, and graph-based information. The hybrid and ensemble models consistently achieved a higher accuracy, robustness, and adaptability compared to the standalone ML and DL models [8,30,32] and [6]. The results are consistent with the review finding that currently hybrid architectures are considered the state of the art for OSN spam detection. However, their architectural complexity, training burden, and deployment expenses are still significant challenges to ensure their use at large scale.

Optimization and robustness techniques were also found to be crucial parts of modern-day spam detection techniques. Various techniques such as GAN-based augmentation, SMOTE-ENN balancing strategies, and creation and optimization of model parameters with Reinforcement Learning and the use of Genetic Algorithms helped improve the performance of the models in generalization, avoiding overfitting, and being resistant to adversarial spam attacks. Several works like [23], [37], [36], [33] and [35] have demonstrated that optimizing learning strategies yielded improved performance. However, the optimization process was often complex and time-consuming, and implementation was difficult when it involved high-dimensional OSN datasets. These are many of the time-old problems in model robustness and computational efficiency. Furthermore, for optimizing spam training without altering the classification architecture, the SON2S/Modified Genetic Algorithm-XGBoost uses intelligent parameter tuning and feature-selection methodology, which demonstrates the criticality of using intelligent parameter tuning and feature-selection techniques to improve the spam detection system.

As for explainability and robustness (RQ3), attention mechanisms and graph-based learning have been employed to increase the interpretability of spam content in some of the research discussed above, but few works have studied the application of explicit methods of Explainable Artificial Intelligence (XAI) like SHAP [13], LIME [14], or interpretable attention visualization for spam detection in existing studies. But most reviewed studies were directed towards performance optimization and not transparency and interpretability. This restriction is of concern for automated content moderation systems used on large social media platforms for trust, accountability, and fairness.

Last, challenges related to the datasets (RQ4) were identified as important. Many of the studies were based on heavily skewed, domain-specific, and outdated benchmark datasets from Twitter or from one of the other platforms: YouTube, Yelp, and Amazon. There are not many datasets that are available on a large scale that are both multilingual and multimodal, which can limit the extent to which the models can be generalized to other OSN platforms and changing anti-spam tactics. Furthermore, the augmentation using synthetic methods (SMOTE, GAN) may not represent the actual adversarial changes occurring in spam in the real world, which could affect the generalizability in a dynamic deployment.

Another key caveat in implementing the results of this review is the variation of the data sets used in the different studies included. These reviewed studies relied on information available on a wide variety of platforms, such as Twitter, YouTube, Amazon, and Yelp, which vary significantly in their size, features, distribution of classes, languages used and those of the users, and patterns of user actions, as well as the type of spam-like patterns. Together, the parameters of the included datasets were organized and summarized in Table 4.

Table 4. Summary of Dataset Characteristics Across Reviewed Studies

Platform / Dataset Type	Number of Studies	Typical Size	Class Balance	Key Spam Type
Twitter	16	10K–500K tweets	Imbalanced (5-30% spam)	Bot accounts, spam tweets and phishing
YouTube	1	~1,900 comments	Moderately balanced	Spam comments
Amazon / Yelp	2	10K–100K reviews	Imbalanced	Fake and spam reviews
Multi-platform / Social network graphs	6	Varies	Varies	Coordinated spam and fraud

The variations in platform characteristics, feature spaces, and class distributions have a significant impact on reported platform performance metrics. However, results comparisons should be made with care between different studies since the gains obtained on one dataset (context) with respect to these metrics do not necessarily transfer to other datasets (context) or to other types of spam. The conclusions of this review should thus be interpreted as

reflecting the trend of methodology in general, and not as a final and certain decision of one methodology over the others in all spam detection contexts. This constraint is also important to note since it underscores the need for developing publicly available, cross-platform, standard benchmark datasets now for the sake of fair comparisons in future research.

The conclusions drawn from this review suggest that the hybrid DL, GNN, and Transformer models give better spam detection performance, but future research should aim to develop models that are scalable, robust, efficient, and interpretable to facilitate the deployment of these models in large-scale and real-time OSNs.

5. Recommendations

The findings of this systematic review led to the following recommendations for future research and in the design and implementation of spam detection systems in OSNs:

- There is a need for future research to develop effective real-time deployable lightweight and scalable hybrid architectures with the incorporation of GNNs, Transformer models, and attention mechanisms that could benefit from each technique.
- Explainable Artificial Intelligence (XAI) should be promoted, e.g., using SHAP, LIME, and interpretable attention mechanisms, increasing transparency, accountability, and user trust in automation for spam detection systems.
- Research is required for advanced multimodal spam detection systems by using a mixture of textual, visual, behavioral, and relational information, resulting from more sophisticated spam campaigns.
- Analyzing future models on different social media platforms can boost generalizability across platforms and reduce the need for platform-specific datasets.
- There is a need for more research on scalable graph processing techniques and distributed learning frameworks that can be used efficiently to process large social network datasets.
- Adversarial learning, continual learning, and robustness-enhancement strategies should be added for powerful resistance to new spam attacks and ensemble operations of bots.
- Multilingual, publicly available, and frequently updated benchmark sets that enable the reproducibility, fair comparison, and generalization of spam detection models should be created by the research community.
- Explore methods like model compression, pruning, knowledge distillation, and efficient attention mechanisms to reduce the computational complexity of Transformer-based and hybrid models.
- New deployments beyond classification performance require a variety of metrics such as inference latency, memory usage, energy efficiency, and scalability to be evaluated.
- There is a need for future OSN spam detection systems to consider ethical concerns regarding privacy, fairness of algorithms, transparency, and responsible use of AI.

6. Conclusion

The present systematic review sought to summarize the knowledge extracted from 25 high-quality studies, which were published from 2020 to 2025 on the research development of spam detection for OSNs. Using Traditional ML, DL, GNNs, Transformer-based approaches, Hybrid Models, and Optimization Techniques, spam, malicious accounts, phishing, and spam/phishing detection were analyzed in social media.

The results show a substantial deviation from classic ML models to more complex models, such as deep learning and graph-based approaches. CNNs and RNNs are effective; however, they are complicated. Traditional ML algorithms such as KNN, NB, DT, and RF still retain their own advantages with ease of use, lower computational costs, interpretability, etc. However, they are not always successful because of new spam-seeder behaviors and complex contexts. In contrast, the performance of automatic feature extraction, contextual understanding, relational learning, and semantic representation was found to be good in several studies, which bolstered the spam detection ability of DL, GNN, and Transformer-based models.

The review concluded that hybrid and ensemble models overall outperformed other classification approaches and used the best features of each model for greatest success on spamming. These were further enhanced with SOA, data augmentation via GANs, hyperparameter optimization, and class-balancing techniques. These benefits have been achieved at the cost of challenging deployment and high computational complexity, memory usage, and training duration, however.

A recognized key limitation was that some studies did not involve the same data and social networks, some feature different representations, some reports have different class distributions, and some have different

evaluation protocols, so that direct quantitative cross-study comparisons are not always easy. The conclusions therefore offer insights into general trends in methodology but are not intended to criticize or advocate any individual methodology. To facilitate future development of scalable spam detection systems addressing detection precision, computational efficiency, robustness, explainability, real-time applicability, and user-friendliness, advanced AI models and techniques in hybrid deep learning, graph-based learning, Transformer architectures, optimization, and Explainable Artificial Intelligence (XAI) should be leveraged.

Funding: This work did not receive any external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

AI Use Declaration: Generative AI was used solely for language editing and proofreading. The authors reviewed and approved all content and take full responsibility for the manuscript.

References:

1. Schmitt, M.; Flechais, I. Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. *Artif. Intell. Rev.* **2024**, *57*, doi:10.1007/s10462-024-10973-2.
2. Mouncey, E.; Ciobotaru, S. Phishing Scams on Social Media: An Evaluation of Cyber Awareness Education on Impact and Effectiveness. *Journal of Economic Criminology* **2025**, *7*, 100125, doi:10.1016/J.JECONC.2025.100125.
3. Simon, M.; Ya'u, B.I.; Gital, A.Y.; Magombe, Y.; Alli, A.; Maishanu, M. A Hybrid Convolutional Neural Network-Gated Recurrent Unit Model for Fake News Detection on Social Media. In Proceedings of the 2026 2nd International Symposium on AI-Driven Engineering Systems (ISADES); IEEE, 2026; pp. 1–5. doi**:** 10.1109/ISADES69945.2026.11608227
4. Pierri, F.; Luceri, L.; Jindal, N.; Ferrara, E. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. *ACM International Conference Proceeding Series* **2023**, 65–74, doi:10.1145/3578503.3583597.
5. Kawintiranon, K.; Singh, L.; Budak, C. Traditional and Context-Specific Spam Detection in Low Resource Settings. *Mach. Learn.* **2022**, *111*, 2515–2536, doi:10.1007/S10994-022-06176-X.
6. Shaaban, M.A.; Hassan, Y.F.; Guirguis, S.K. Deep Convolutional Forest: A Dynamic Deep Ensemble Approach for Spam Detection in Text. *Complex and Intelligent Systems* **2022**, *8*, 4897–4909, doi:10.1007/s40747-022-00741-6.
7. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning-Based Text Classification. *ACM Comput. Surv.* **2022**, *54*, https://doi.org/10.1145/3439726
8. Rao, S.; Verma, A.K.; Bhatia, T. Hybrid Ensemble Framework with Self-Attention Mechanism for Social Spam Detection on Imbalanced Data. *Expert Syst. Appl.* **2023**, *217*, 119594, doi:10.1016/J.ESWA.2023.119594.
9. Zhou, M.; Zhang, D.; Wang, Y.; Geng, Y.A.; Dong, Y.; Tang, J. LGB: Language Model and Graph Neural Network-Driven Social Bot Detection. *IEEE Trans. Knowl. Data Eng.* **2024**, *37*, 4728–4742, doi:10.1109/TKDE.2025.3573748.
10. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, 790–799, doi:10.1136/BMJ.N71.
11. Wang, K.; Ding, Y.; Han, S.C. Graph Neural Networks for Text Classification: A Survey. *Artif. Intell. Rev.* **2024**, *57*, doi:10.1007/s10462-024-10808-0.
12. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks. *Expert Syst. Appl.* **2023**, *241*, doi:10.1016/j.eswa.2023.122666.
13. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence 2020 2:1* **2020**, *2*, 56–67, doi:10.1038/s42256-019-0138-9.
14. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should i Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, *13-17-August-2016*, 1135–1144, https://doi.org/10.1145/2939672.2939778
15. Bhattacharya, M.; Roy, S.; Chattopadhyay, S.; Das, A.K.; Shetty, S. A Comprehensive Survey on Online Social Networks Security and Privacy Issues: Threats, Machine Learning-Based Solutions, and Open Challenges. *Security and Privacy* **2023**, *6*, e275, doi:10.1002/SPY2.275.
16. Xiao, A.S.; Liang, Q. Spam Detection for Youtube Video Comments Using Machine Learning Approaches. *Machine Learning with Applications* **2024**, *16*, 100550, doi:10.1016/J.MLWA.2024.100550.
17. Gangwar, S.S.; Rathore, S.S.; Chouhan, S.S.; Soni, S. Predictive Modeling for Suspicious Content Identification on Twitter. *Soc. Netw. Anal. Min.* **2022**, *12*, 149, doi:10.1007/S13278-022-00977-7.
18. Velammal, B.L.; Aarthy, N. Improvised Spam Detection in Twitter Data Using Lightweight Detectors and Classifiers. *International Journal of Web-Based Learning and Teaching Technologies* **2021**, *16*, 12–32, doi:10.4018/IJWLTT.20210701.OA2.
19. Alharthi, R.; Alhothali, A.; Moria, K. A Real-Time Deep-Learning Approach for Filtering Arabic Low-Quality Content and Accounts on Twitter. *Inf. Syst.* **2021**, *99*, 101740, doi:10.1016/J.IS.2021.101740.
20. Hayami, R.; Al Amien, J.; Utami, N.T. Tweet Spam Detection Using the Convolutional Neural Network (CNN) Model. *AIP Conf. Proc.* **2023**, *2601*, doi:10.1063/5.0130468/2894141.
21. Loucif, H. A Hybrid Deep Learning Approach for Spam Detection in Twitter. *Ingenierie des Systemes d'Information* **2024**, *29*, 117–123, doi:10.18280/ISI.290113.
22. Xu, G.; Zhou, D.; Liu, J. Social Network Spam Detection Based on ALBERT and Combination of Bi-LSTM with Self-Attention. *Security and Communication Networks* **2021**, *2021*, 5567991, doi:10.1155/2021/5567991.

23. Kumar, C.; Bharti, T.S.; Prakash, S. A Hybrid Data-Driven Framework for Spam Detection in Online Social Network. *Procedia Comput. Sci.* **2023**, *218*, 124–132, doi:10.1016/J.PROCS.2022.12.408.
24. Pal, M.B.; Agrawal, S. Graph Neural Network-Based Attention Mechanism to Classify Spam Review over Heterogeneous Social Networks. *J. Supercomput.* **2024**, *80*, 27176–27203, doi:10.1007/S11227-024-06459-1.
25. Dou, Y.; Liu, Z.; Sun, L.; Deng, Y.; Peng, H.; Yu, P.S. Enhancing Graph Neural Network-Based Fraud Detectors against Camouflaged Fraudsters. *International Conference on Information and Knowledge Management, Proceedings* **2020**, 315–324, doi:10.1145/3340531.3411903.
26. Zhao, C.; Xin, Y.; Li, X.; Zhu, H.; Yang, Y.; Chen, Y. An Attention-Based Graph Neural Network for Spam Bot Detection in Social Networks. *Applied Sciences 2020, Vol. 10, Page 8160* **2020**, *10*, 8160, doi:10.3390/APP10228160.
27. Shen, H.; Liu, X.; Zhang, X. Boosting Social Spam Detection via Attention Mechanisms on Twitter. *Electronics 2022, Vol. 11, Page 1129* **2022**, *11*, 1129, doi:10.3390/ELECTRONICS11071129.
28. Sallah, A.; Arbi Abdellaoui Alaoui, E.; Agoujil, S.; Ahmad Wani, M.; Hammad, M.; Maleh, Y.; Abd El-Latif, A.A. Fine-Tuned Understanding: Enhancing Social Bot Detection With Transformer-Based Classification. *IEEE Access* **2024**, *12*, 118250–118269, doi:10.1109/ACCESS.2024.3440657.
29. Nasser, M.; Saeed, F.; Da'u, A.; Alblwi, A.; Al-Sarem, M. Topic-Aware Neural Attention Network for Malicious Social Media Spam Detection. *Alexandria Engineering Journal* **2025**, *111*, 540–554, doi:10.1016/J.AEJ.2024.10.073.
30. Iqbal, A.; Younas, M.; Iftikhar, S.; Fatima, F.; Saleem, R. Spam Detection Using Hybrid Model on Fusion of Spammer Behavior and Linguistics Features. *Egyptian Informatics Journal* **2025**, *29*, 100605, doi:10.1016/J.EIJ.2024.100605.
31. Citlak, O.; Dorterler, M.; Dogru, İ. A Hybrid Spam Detection Framework for Social Networks. *Politeknik Dergisi* **2023**, *26*, 823–837, doi:10.2339/POLITEKNIK.933785.
32. Bilgen, Y.; Kaya, M. EGMA: Ensemble Learning-Based Hybrid Model Approach for Spam Detection. *Applied Sciences 2024, Vol. 14, Page 9669* **2024**, *14*, 9669, doi:10.3390/APP14219669.
33. Elakkiya, E.; Selvakumar, S. Stratified Hyperparameters Optimization of Feed-Forward Neural Network for Social Network Spam Detection (SON2S). *Soft Computing 2022 26:21* **2022**, *26*, 11915–11934, doi:10.1007/S00500-022-07020-Z.
34. Abualhaj, M.M.; Daoud, M.S.; Al-Allawee, A.; Al-Mimi, H.; Al-Khatib, S.; Hyassat, A.S. Detecting Spam Using K-Nearest Neighbors and Metaheuristic Feature Selection Algorithms. *International Conference on Social Networks Analysis, Management and Security, SNAMS* **2025**, 249–252, doi:10.1109/SNAMS67467.2025.11390963.
35. Shukla, P.K.; Veerasamy, B.D.; Alduaiji, N.; Addula, S.R.; Pandey, A.; Shukla, P.K. Fraudulent Account Detection in Social Media Using Hybrid Deep Transformer Model and Hyperparameter Optimization. *Scientific Reports 2025 15:1* **2025**, *15*, 38447-, doi:10.1038/s41598-025-24326-8.
36. Ghatasheh, N.; Altaharwa, I.; Aldebei, K. Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction. *IEEE Access* **2022**, *10*, 84365–84383, doi:10.1109/ACCESS.2022.3196905.
37. Rashidi, A.; Salehi, M.; Najari, S. CGANS: A Code-Based GAN for Spam Detection in Social Media. *Social Network Analysis and Mining 2024 14:1* **2024**, *14*, 218-, doi:10.1007/S13278-024-01379-7.
38. Chandana; Anuradha; Naik, A.J.; Kumar, A.; Banu, S. A Framework for Twitter Spam Detection and Reporting. *AIP Conf. Proc.* **2024**, *2742*, doi:10.1063/5.0184161/3263654.

Appendix A. Data Extraction Form

The data extraction form presented in Table A1 was developed before the review process and was used to systematically collect and organize the relevant information from each study included in the systematic review. The extracted information enabled consistent comparison of methodologies, datasets, evaluation metrics, and key findings across all selected studies.

Table A1. Data Extraction Form

Extraction Field	Description
Author(s) and Year	Name(s) of author(s) and publication year
Study Objective	Purpose of the research
Detection Approach	ML, DL, GNN, Transformer, Hybrid, or Ensemble method
Dataset Used	Dataset name, source, and platform (Twitter, SMS, Facebook, etc.)
Dataset Characteristics	Size, class balance, public availability
Feature Types	Textual, behavioral, relational, contextual, or multimodal
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, AUC
Computational Efficiency	Training time, memory usage, scalability
Explainability Technique	SHAP, LIME, Attention, XAI methods
Robustness Method	Adversarial learning, imbalance handling, optimization
Main Findings	Key results and contributions
Limitations	Weaknesses and challenges identified
Future Recommendation	Suggested future improvements

Appendix B. Summary of Reviewed Studies

Appendix B summarizes the 25 studies included in the final systematic review. Table B1 presents the principal characteristics of each study, including its research focus, methodological approach, and the dataset or platform used for evaluation. This summary provides an overview of the diversity of spam detection approaches investigated in the reviewed literature.

Table B1. Summary of Reviewed Studies

S/N	Author(s) & Year	Study Focus	Methods/Models Used	Dataset/Platform
1	[25].	Fraud and spam detection in OSNs	Graph Convolutional Network (GCN)	Social network graph datasets
2	[26]	Spam bot detection in social networks	Graph Attention Network (GAT)	Social network datasets
3	[19]	Twitter spam detection	CNN-LSTM hybrid model	Twitter
4	[22]	Social network spam detection	ALBERT + Bi-LSTM with Self-Attention	Social network spam dataset
5	[28]	Social bot detection in online social networks	BERT, GPT variants, Feedforward Neural Network (FNN), Explainable AI (XAI)	Twitter bot dataset
6	[18]	Twitter spam tweet classification	Naïve Bayes, Logistic Regression, Random Forest	Twitter
7	[17]	Suspicious Twitter content detection	DT, RF, KNN, LR, NB with hybrid features	Twitter
8	[27]	Attention-driven spam detection	BERTweet + Attention + GAT	Twitter
9	[6]	Dynamic deep ensemble spam detection	CNN + RF + Extremely Randomized Trees	Text spam datasets
10	[33]	Social network spam detection	SON2S, FFNN, RL, Modified SFLA	Twitter Dataset, Tip Spam Dataset
11	[20]	Tweet spam detection	CNN-based tweet classification	Twitter
12	[23]	Spam detection on imbalanced datasets	SMOTE-ENN + Deep Learning	Twitter
13	[31]	Social network spam detection	Hybrid ML + Text + Short-link analysis	Social networks
14	[8]	Social spam detection on imbalanced data	Hybrid Ensemble + Self-Attention	Social spam datasets
15	[38]	Twitter spam classification	RF, KNN, NB, DT	Twitter
16	[24]	Spam review classification	Attention-based Graph Neural Network	Amazon, YelpChi
17	[32]	Ensemble spam detection framework	EGMA (GRU + Autoencoder + MLP)	Multiple spam datasets
18	[36]	Spam prediction and detection in social networks	Modified Genetic Algorithm (GA), extreme Gradient Boosting (XGBoost), Feature Selection and Hyperparameter Optimization	Twitter, SMS Spam
19	[37]	GAN-based social spam detection	Conditional GANs (CGANS)	Social media datasets
20	[16]	YouTube spam detection	RF, DT, KNN, SVM, LR, NB	YouTube comments
21	[21]	Twitter spam detection	Hybrid CNN–LSTM + WordNet + ConceptNet	Twitter
22	[35]	Fraudulent account detection	TCN + GAN + SOA + Autoencoder	Twitter/TwiBot datasets
23	[29]	Malicious social spam detection	LSTM + Topic Modeling + BERT + Attention	Social media datasets
24	[34]	Spam and malicious URL detection	KNN + Metaheuristic Feature Selection	Social network datasets
25	[30]	Behavioral and linguistic spam detection	Hybrid behavioral linguistic fusion model	Social review datasets

The 25 main studies that were analyzed in this systematic review are summarized in Appendix B and classified based on their selected spam detection paradigm, solution method, and testing platform.

Appendix C. Details of Included Studies

To improve transparency and reproducibility, Table C1 lists the complete bibliographic titles of all studies included in the final systematic review. This appendix allows readers to clearly identify the reviewed publications referred to throughout the manuscript.

Table C1. Complete List of Included Studies

S/N	Author(s) & Year	Paper Title
1	[25].	Enhancing Graph Neural Network based Fraud Detectors against Camouflaged Fraudsters
2	[26]	An Attention-Based Graph Neural Network for Spam Bot Detection in Social Networks
3	[19]	Deep Learning Approaches for Social Spam Detection on Twitter Using CNN-LSTM Architectures
4	[22]	Social Network Spam Detection Based on ALBERT and Combination of Bi-LSTM with Self Attention
5	[28]	Fine-Tuned Understanding: Enhancing Social Bot Detection with Transformer-Based Classification
6	[18].	Improvised Spam Detection in Twitter Data Using Lightweight Detectors and Classifiers
7	[17]	Predictive Modeling for Suspicious Content Identification on Twitter
8	[27]	Boosting Social Spam Detection via Attention Mechanisms on Twitter
9	[6]	Deep Convolutional Forest: A Dynamic Deep Ensemble Approach for Spam Detection in Text
10	[33]	Stratified hyperparameters optimization of feed-forward neural network for social network spam detection (SON2S)
11	[20]	Tweet Spam Detection Using the Convolutional Neural Network (CNN) Model
12	[23]	A Hybrid Data-Driven Framework for Spam Detection in Online Social Network
13	[31]	A Hybrid Spam Detection Framework for Social Networks
14	[8]	Hybrid Ensemble Framework with Self-Attention Mechanism for Social Spam Detection on Imbalanced Data
15	[38]	Twitter Spam Detection Using Machine Learning Classifiers
16	[24]	Graph Neural Network-Based Attention Mechanism to Classify Spam Review over Heterogeneous Social Networks
17	[32]	EGMA: Ensemble Learning-Based Hybrid Model Approach for Spam Detection
18	[36]	Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction
19	[37]	CGANS: A Code-Based GAN for Spam Detection in social media
20	[16]	Spam Detection for Youtube Video Comments Using Machine Learning Approaches
21	[21]	A Hybrid Deep Learning Approach for Spam Detection in Twitter
22	[35]	Fraudulent Account Detection in social media Using Hybrid Deep Transformer Model and Hyperparameter Optimization
23	[29]	Topic-Aware Neural Attention Network for Malicious Social Media Spam Detection
24	[34]	Detecting Spam Using K-Nearest Neighbors and Metaheuristic Feature Selection Algorithms
25	[30]	Spam Detection Using Hybrid Model on Fusion of Spammer Behavior and Linguistics Features

Appendix D. Quality Assessment Results

The methodological quality of the eligible studies was evaluated using a predefined quality assessment framework adapted from established guidelines for conducting systematic reviews in computer science. Following title, abstract, and full-text screening, 84 studies met the eligibility criteria for quality assessment, of which 25 studies satisfied the required methodological quality and were included in the final synthesis.

Each study was evaluated using six quality assessment criteria:

- **DS** – Data Suitability
- **MT** – Methodological Transparency
- **ER** – Evaluation Robustness
- **IC** – Innovation and Contribution
- **EX** – Explainability
- **SC** – Scalability Consideration

Each criterion was scored as follows:

- **2** = Fully satisfied
- **1** = Partially satisfied
- **0** = Not satisfied

The maximum possible quality score for each study was 12 points.

The studies were classified according to the following quality levels:

- **10–12:** High Quality
- **7–9:** Moderate Quality
- **0–6:** Low Quality

Only studies rated as Moderate or High Quality were included in the final systematic synthesis.

Table D1. Quality Assessment Scores of Included Studies

Study	DS	MT	ER	IC	EX	SC	Total	Quality
[25].	2	2	2	2	1	1	10	High
[26]	2	2	2	2	1	1	10	High
[19]	2	2	2	2	0	1	9	Moderate
[22]	2	2	2	2	1	1	10	High
[28]	2	2	2	2	2	1	11	High
[18].	2	2	2	1	0	1	8	Moderate
[17]	2	2	2	1	0	1	8	Moderate
[27]	2	2	2	2	1	1	10	High
[6]	2	2	2	2	0	1	9	Moderate
[33]	2	2	2	2	1	1	10	High
[20]	2	2	2	1	0	1	8	Moderate
[23]	2	2	2	2	0	1	9	Moderate
[31]	2	2	2	2	0	1	9	Moderate
[8]	2	2	2	2	1	1	10	High
[38]	2	2	2	1	0	1	8	Moderate
[24]	2	2	2	2	1	1	10	High
[32]	2	2	2	2	0	1	9	Moderate
[36]	2	2	2	2	1	1	10	High
[37]	2	2	2	2	0	1	9	Moderate
[16]	2	2	2	1	0	1	8	Moderate
[21]	2	2	2	2	0	1	9	Moderate
[35]	2	2	2	2	1	1	10	High
[29]	2	2	2	2	1	1	10	High
[34]	2	2	2	2	0	1	9	Moderate
[30]	2	2	2	2	1	1	10	High

Table D2. Summary of Quality Assessment Statistics

Quality Category	Number of Studies	Percentage
High Quality (10-12)	12	48.0%
Moderate Quality (7-9)	13	52.0%
Low Quality (0-6)	0	0.0%
Total Included Studies	25	100.0%

Overall, 12 studies (48.0%) were classified as High Quality, while 13 studies (52.0%) were classified as Moderate Quality. No studies included in the final synthesis were rated as Low Quality, indicating that all selected studies satisfied the predefined methodological quality requirements.