electronics

MDPI

Systematic Review

Advances in NLP Techniques for Detection of Message-Based Threats in Digital Platforms: A Systematic Review

José Saias [1,2]

[1] Departamento de Informática, Escola de Ciências e Tecnologia, Universidade de Évora, Rua Romão Ramalho, n. 59, 7000-671 Évora, Portugal; jsaias@uevora.pt
[2] VISTA Lab, ALGORITMI Research Centre/LASI, University of Évora, 7004-516 Évora, Portugal

Abstract

Users of all ages face risks on social media and messaging platforms. When encountering suspicious messages, legitimate concerns arise about a sender's malicious intent. This study examines recent advances in Natural Language Processing for detecting message-based threats in digital communication. We conducted a systematic review following PRISMA guidelines, to address four research questions. After applying a rigorous search and screening pipeline, 30 publications were selected for analysis. Our work assessed the NLP techniques and evaluation methods employed in recent threat detection research, revealing that large language models appear in only 20% of the reviewed works. We further categorized detection input scopes and discussed ethical and privacy implications. The results show that AI ethical aspects are not systematically addressed in the reviewed scientific literature.

Keywords: NLP; threat detection; cybersecurity; social media; AI ethics

1. Introduction

Both social media and instant messaging platforms pose risks of online predator activity, catfishing, dating scams, and social engineering attacks aimed at data harvesting, including phishing and spear phishing, as well as the spread of malware. In online messaging and social platforms, when a suspicious message is received, several critical questions arise:

a Has the account of the person I am interacting with been compromised?
b Is this unknown sender a real person acting with malicious intent?
c Could this be part of a manipulation scam or a targeted social engineering attempt?
d Is the sender impersonating someone else?
e Could that profile be fake and controlled by a bot for deceptive purposes?

A variety of studies have explored these topics. In [1], the authors explore how the convergence of social media platforms and Artificial Intelligence (AI) foster the possibility of exploitation, using case studies to examine victims' experiences in digital romance fraud. In [2], the authors analyze romance fraud and behavioral changes on the Internet use during the pandemic. The challenges of fake profiles and the techniques used to detect them were surveyed in [3]. The dual role of large language models (LLMs) in social engineering is examined in [4]. Describing a case study on collusion scams, the authors analyze both the potential of LLMs to generate attacks and their possible use in enhancing detection. The work in [5] presents a review of deep learning approaches for phishing detection in E-mail.

check for updates

Academic Editor: Arkaitz Zubiaga

Received: 21 April 2025
Revised: 18 June 2025
Accepted: 20 June 2025
Published: 24 June 2025

Citation: Saias, J. Advances in NLP Techniques for Detection of Message-Based Threats in Digital Platforms: A Systematic Review. *Electronics* **2025**, *14*, 2551. https://doi.org/10.3390/electronics14132551

While other studies have focused on surveys of specific types of threats, there is a gap in the literature regarding analyses from a more transversal perspective, one that emphasizes common patterns, particularly the threats' action channel (messages) and medium (text). Additionally, the practical applicability of these strategies in a real-world context, especially in light of emerging ethical and legal challenges, remains underexplored.

Our work focuses on recent advancements in Natural Language Processing (NLP) techniques for identifying unsafe conversations in private or direct messaging environments. This review is guided by the following research questions:

RQ1. What NLP techniques are employed in threat detection for messaging platforms, and how are LLMs integrated into these approaches, according to recent research?
RQ2. What evaluation strategies are used to measure the effectiveness of these techniques?
RQ3. What types of input are used for online threat detection in recent studies?
RQ4. How can detection efficacy be balanced with privacy, informed consent, and compliance with legal and ethical regulations?

RQ1 directly targets the central research gap concerning a holistic perspective on textual and message-based threats. Rather than focusing on isolated threat categories, it investigates NLP techniques underpinning diverse forms of online deception. Given the paradigm-shifting impact of LLMs, we explicitly incorporate their role in our analysis. The second research question seeks to evaluate the performance of detection methods, and to analyze how recent studies assess their effectiveness. In online security contexts, where missed threats may have severe consequences, identifying potential limitations in current approaches helps to prioritize future research directions. While messages serve as the primary threat vector, effective detection typically requires the analysis of broader contextual cues. RQ3 characterizes the supplementary information researchers use alongside textual content to enhance detection, enabling a holistic understanding of detection strategies. RQ4 addresses the second identified gap, seeking to understand the suitability of these detection strategies for real use in citizen protection, and to verify whether recent studies explicitly account for user privacy and regulatory compliance.

According to the *Internet Organised Crime Threat Assessment 2021* report from EuroPol [6], that year, phishing and social engineering increased to generate considerable criminal proceeds. The pandemic-driven shift to online shopping and digital interaction created heightened opportunities for fraud, as malicious actors exploited increased reliance on virtual platforms. Three years after, the *IOCTA2024* report [7], highlighted the growing use of end-to-end encrypted communication platforms by offenders. Phishing remained the leading fraud method in 2023 against EU citizens, companies, and institutions, and Smishing was the most commonly used variant. The North-American *Internet Crime Report 2023* (https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf (accessed on 2 April 2025)), accounts for 51,750 complaints on impersonation scams, with total losses surpassing USD 1.3 billion that year.

Securing online interactions is critical for individuals and organizations alike, as compromised personal data can expose affiliated entities to targeted attacks. With these assumptions, the main contributions of this work are the following:

1. A characterization of message-based threat detection strategies as reported in the recent scientific literature and the extent to which LLMs are involved;
2. Understand whether there is standardization or not of the evaluation strategies of contemporary approaches;
3. Determine the monitoring scope selected in research for threat detection;
4. How current research addresses new challenges regarding ethics and privacy in the use of data in a real-world scenario.

The following article content has been structured as follows. Section 2 describes the context and the main types of messaging-based threats as well as the current AI and NLP landscape. Section 3 describes the review search strategy, the inclusion criteria, and the data extraction process. In Section 4, we present the outcomes of each phase of our systematic review protocol. We discuss the results and the interpretations that can be drawn from them in Section 5. Finally, in Section 6, we summarize our main findings and present future research directions.

2. Problem Context

This section presents foundational concepts and essential terminology relevant to this research area and mentioned in later sections.

2.1. Message-Based Threats

Some threats associated with message-based interaction have been known for decades, but they have evolved with technological advances, creating the need for continuous updates to protection techniques. When replying to a message (or reacting to it in certain ways), certain risks arise that are independent of the terminal device, whether a computer or smartphone, and are unrelated to the specific software used, such as browsers or apps. These risks can lead to both short-term and long-term harm to victims. Sometimes the consequence only occurs after a period in which the attacker gradually builds trust and gathers valuable information. Some of the types of threats conveyed through messages are as follows:

- Fake profile is a fake online persona designed to mislead or deceive other users. It may involve photos, names, or details copied from other people, or artificially fabricated by AI. It can be operated by a bot, controlled by AI, or directly by a human.
- Impersonation is the malicious use of another person's identity, by creating a fake profile, or by hacking into an existing legitimate account. By impersonating a celebrity, for example, the attacker may gain advantages and convince victims to provide information while thinking they are actually talking to that person.
- Shaming and cancel culture is the act of explicit criticism towards victims, through comments, group chats, or posts online, either for offensive reasons or in an attempt to shape public opinion against them.
- Pretexing is a form of social engineering in which the victim is confronted with an unreal and unexpected scenario (the pretext), and the attacker tries to manipulate them with persuasive storytelling.
- Social engineering is a manipulation technique that seeks to take advantage of human behavior and psychology and is used to deceive individuals into revealing valuable information, performing actions, or instinctively react in a compromising way.
- Phishing is a form of cyberattack where criminals use forged messages or web content to scam victims into disclosing sensitive information. Usually it comprises two parts: the bait, or deceptive content alluding to some trusted entity, and the hook, or a way of capturing information or inserting malware.
- Spear phishing is a form of phishing targeted at a specific individual or organization, often involving prior research to include credible and personalized elements.
- Smishing is a phishing variant where the attack is conducted via text messages (SMS). The name first S stands for SMS, just like in Vishing, where the V stands for voice call-based phishing.
- Doxing (also referred to as 'doxxing') is the malicious act of exposing someone's private or sensitive information in public or within a group, in an unauthorized manner, and with the purpose of harassing or humiliating.

- Spam is the process of sending unsolicited (or simply irrelevant) messages, usually to a large number of recipients. It can be used to send advertising or to spread rumors or misinformation, but it can also be a channel for phishing.
- A scam is a broad class of deliberate fraudulent schemes designed to manipulate victims into losing money, or to obtain control, resources, or valuable information.
- Harmful content, as a threat, is the transmission of material that inflicts emotional, physical, or societal harm or was created through any form of abuse or exploitation.
- Hate speech is a form of harmful content that promotes hostility against individuals or groups.

It can be observed that some concepts are specializations of broader categories. In addition to this hierarchical structure, some threats may span across multiple categories.

Terminological ambiguity is occasionally observed. Some studies refer to "fake account", although "fake profile" is generally the more accurate term. While an account may be compromised, it technically refers to a technical concept linked to platform access, and is not fake in that sense but rather in terms of the false identity it is intended to represent.

2.2. Landscape of NLP and AI

This section examines the fundamental concepts of NLP and AI, with a focus on pertinent techniques and recent developments in these evolving domains.

2.2.1. NLP

Natural Language Processing (NLP) is a branch of computer science that blends Artificial Intelligence with Computational Linguistics, empowering computers to interpret, analyze, and produce human language. Many applications benefit from NLP techniques, including named entity recognition, sentiment analysis, question-answering, dialogue systems, machine translation, speech recognition, and text-to-speech technology [8].

Typical NLP processing pipelines involve preprocessing, feature extraction to convert raw text into representations interpretable by machine learning models, and model training. Some NLP techniques do not rely on ML and often involve rule and heuristic-based approaches and statistical approaches, which do not require training.

Since around 2010, deep neural networks have significantly improved the performance of many NLP tasks, driving major advances in the field.

2.2.2. ML and DL

Machine learning (ML) is a branch of Artificial Intelligence focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data [9]. Some traditional models—meaning not based on neural networks—rely on manual features or statistical patterns. Some algorithms falling into this class are Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF).

Deep learning (DL) is a class of ML techniques that use multiple layers of nonlinear processing units in neural networks (or deep neural networks) to model complex patterns in data [10]. Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM, a type of RNN) networks are all commonly applied DL algorithms.

Introduced in 2017, Transformers are a type of DL models particularly well suited for tasks involving sequential data, such as language, that leverage a mechanism called self-attention, to process the entire sequence of data in parallel, which leads to significant improvements in efficiency and performance [11]. Bidirectional Encoder Representations

from Transformers (BERT), Bidirectional and Auto-Regressive Transformers (BART), are examples of transformer-based models used in NLP tasks.

2.2.3. LLMs

A large language model (LLM) is a type of language model trained on extremely large volumes of data, having billions of parameters, and aiming to understand and to generate human-like text. LLMs are considered a subset of Transformer-based models and are used to perform several NLP tasks. Between 2019 and 2020, Generative Pre-trained Transformers (GPTs) emerged, in particular, those by OpenAI (https://openai.com/ (accessed on 2 April 2025)), whose popularity grew exponentially with the conversational ChatGPT in 2022. The models were closed and accessible via API. Closed LLMs are proprietary AI models whose weights and source code are not publicly accessible, and interaction typically occurs through an API. On the other hand, open models make their weights available, allowing for inspection or fine-tuning, though they may still have usage restrictions. Going a step further, open-source models provide full access to their training data, code, and documentation, giving users complete control over the model's internals [12].

Some entities and platforms have contributed to facilitating collaboration and access to language models. Hugging Face (https://huggingface.co/ (accessed on 2 April 2025)) is a company and a collaborative online platform that has emerged as an important hub for the machine learning community in accessing models and datasets.

Later, high-quality models started to appear available to researchers, especially after the release of the first LLaMA (https://www.llama.com/ (accessed on 2 April 2025)) model, in February 2023. Meta released LLaMA 2 in July 2023, making it fully accessible to researchers, developers, and businesses. During 2024, Meta followed up with the release of LLaMA 3, while Google also introduced its Gemma family (https://ai.google.dev/gemma (accessed on 2 April 2025)) of open-source models. Mistral AI (https://mistral.ai/ (accessed on 2 April 2025)) has also established itself as a significant contributor to the open-source LLM community, with models known for their efficiency and performance.

The LLM ecosystem has been very dynamic, with growing interest from the community and several releases of new models of various sizes and tuning specializations. In late 2024 and 2025, new models arrived with improvements in dimension (number of parameters) or in improving multimodal capacity for text, audio, and visual processing.

By revisiting these key temporal references in the adoption of LLMs, we provide a rationale for the time horizon selected for our study, as outlined in the following section.

3. Methodology

This research follows a qualitative approach to gain deeper insight into the strategies proposed in the recent literature, which is a common method for studying problems using non-numerical data [13].

3.1. Search Strategy

For transparency and to reduce the risk of reporting bias, we followed the PRISMA 2020 (https://www.prisma-statement.org/ (accessed on 4 April 2025)) guidelines to organize the review process, including the stages of literature search, study screening, and data extraction.

3.1.1. Data Sources

A comprehensive search was prepared across a set of curated academic databases known for indexing peer-reviewed, high-impact, and domain-relevant literature, and supporting structured and reproducible search. Table 1 lists the paper source databases

chosen for the research review, with the corresponding URL address for each one, from which the search process can be replicated.

Table 1. Paper source databases used in this systematic review.

Database Name	URL
ACM Digital Library	https://dl.acm.org/ (accessed on 4 April 2025)
IEEE Xplore	https://ieeexplore.ieee.org/ (accessed on 4 April 2025)
PubMed	https://pubmed.ncbi.nlm.nih.gov/ (accessed on 4 April 2025)
Scopus	https://www.scopus.com/ (accessed on 4 April 2025)
Web of Science	https://www.webofscience.com/ (accessed on 4 April 2025)

3.1.2. Selection Criteria

The selection criteria were guided by the research questions outlined in Section 1. Given RQ1's focus on LLMs, we included only publications dated 1 January 2024 or later, as this represents the earliest date for research including open LLM models to appear in the literature.

We aimed to refine the search strategy to cover the core themes of the review, namely, NLP, threat detection, and messaging platforms. Through preliminary testing, we found it necessary to broaden the search lexicon by including alternative terms commonly used in the field to describe the task, specific threats, or detection contexts. The search terms used to retrieve publication records from the five selected databases are listed in Figure 1.

```
("natural language processing" OR "language processing" OR "text analysis"
    OR "computational linguistics"
    OR NLP OR LLM OR "large language model" OR "language model" )
AND (
       ("threat detection" OR "malicious intent" OR "cyber threat" OR "abuse detection"
                OR "harmful" OR "spam detection" OR "social engineering" OR "phishing detection"
                OR "fraud detection" OR "scam detection")
        OR ( (compromised OR taken OR fake OR catfish OR honeytrap OR impersonation)
             AND (account OR profile))
             AND (detection OR identification OR classification
           )
    )
AND (
    "messaging" OR "social media"
     OR "social network" OR "chat" OR SMS OR MMS OR "online conversation"
     OR "communication platform" OR "digital communication"
     OR "text message" OR "direct message"
     OR "private message" OR WhatsApp OR Telegram OR Messenger
     OR Discord OR Signal OR Facebook OR Instagram  )
```

Figure 1. The search terms used in the chosen databases to find publications.

The search query was applied to both title and abstract fields when possible or limited to abstracts depending on database interface constraints. As part of the screening protocol, we also established specific criteria for the inclusion and exclusion of studies.
Inclusion criteria:

- Relevance to the research questions;
- Conference or journal articles;
- Indexed in at least one of the selected databases (Section 3.1.1);
- Published within the time frame defined for this review;
- Written in English.

Exclusion criteria:

- The publication date falls outside the time frame defined for this review;
- The report is a pre-print, review, or meta-analysis paper;
- After abstract or full-text screening, a study is excluded when outside the scope of the review and not aligned with the research questions;

- The document does not allow retrieval of at least half of the analysis items specified in Section 3.2;
- The full text of the document is not accessible through academic channels and is restricted behind a paywall;
- The document is inaccessible due to legal or technical constraints.

3.2. Data Extraction

A paper analysis protocol and a data extraction table were developed to standardize the retrieval of key characteristics from the selected studies. Sixteen aspects were established and verified in each work:

1. Targeted threats: the specific types of malicious intent or threats that the research aims to detect.
2. Approach: general type of process, such as classification on text or image, sentiment analysis, regression, or other.
3. Model architecture: the type and architecture of the AI model used for threat detection (ML, DL, Transformer-based, LLM, rule-based or other).
4. Preprocessing: the transformation steps applied to raw data.
5. Text representation: how the text is converted into a model appropriate input, such as Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), n-grams, word embeddings, or contextualized embeddings.
6. Detection features: type of features considered in threat detection, text-based, or not (as behavioral features or network connections-based data).
7. Supported language: the language(s) for which the paper's detection techniques were explicitly designed or evaluated.
8. Input scope: the type of data sources analyzed for threat detection.
9. Evaluation: the assessment form and obtained results.
10. Limitations: the researched approach caveats as identified by the authors of the paper.
11. Future directions: the suggestions for improvements or research next steps proposed by the authors.
12. Deployment platform: the specific messaging platforms or environments where the threat detection techniques are applied.
13. Datasets: description of the datasets used for training and evaluating.
14. Real time: whether the research proposal is feasible for real-time detection.
15. Implementation tools: tools, libraries, or frameworks used for implementation or evaluation.
16. Ethical aspects of real-world deployment: considerations regarding ethical implications of deploying the proposed detection methods, including privacy handling, informed consent, or legal compliance to GDPR, CCPA, or other.

4. Results

Following the methodology outlined in the previous section, this part details the results derived from each step and includes an analysis of the collected data points from each study, addressed according to the research questions.

4.1. Search and Screening

The process of literature search and retrieval was conducted in April 2025. Figure 2 illustrates the PRISMA 2020 flow diagram used in this review to depict the selection process of the relevant literature. In the current context, a study refers to a distinct piece of research, while a report refers to any publication or document that describes that study.

The PRISMA diagram reflects both numbers. In this survey, only one document per study was considered.

The search query was not always correctly interpreted by the databases' search engines, particularly regarding the Boolean AND operator. For this reason, we subsequently applied a local filtering tool to the list of results exported from each database, using the title, abstract, and keywords. This process led to the exclusion of 11,814 records. A total of 78 papers were not selected, as they did not fall within the review's defined time period, the entire year of 2024 and the first quarter of 2025. Selected records and retrieved study reports were screened by a human reviewer. Three papers were not retrieved due to unavailability through regular academic sources. Among the reports assessed for eligibility, 31 papers were excluded based on relevance-related exclusion criteria. As an example, the paper in [14] was filtered out by the second and the fourth exclusion criteria mentioned in Section 3.1.2.

After screening and applying the inclusion criteria, a total of 30 publications were selected for detailed analysis, comprising 11 peer-reviewed journal articles and 19 papers from conference proceedings.

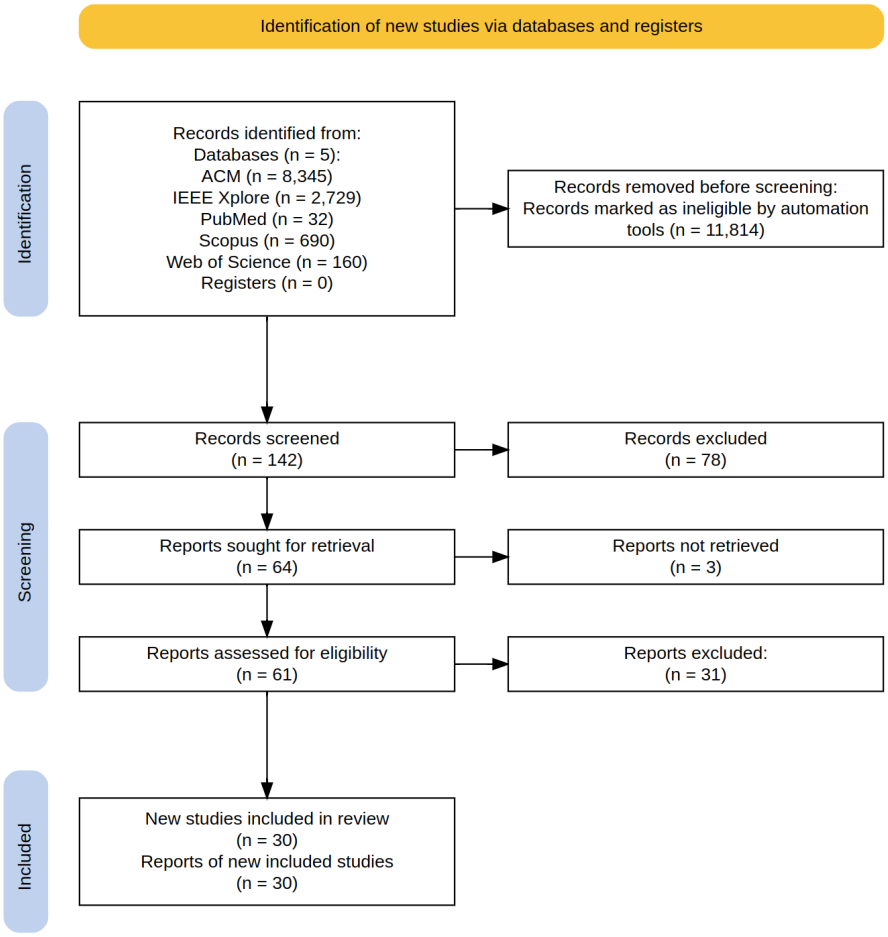

Figure 2. Flow of information through the different phases (PRISMA diagram [15]).

The data items retrieved from the detailed paper analysis (outlined in Section 3.2) are listed across Tables A1–A3, presented in Appendix A.

4.2. Analysis of the NLP Techniques Employed

Beginning with the distribution of studies by threat type, we found five papers focused on phishing [16–20] and three more specifically addressing smishing [21–23], two dealing

with social engineering [24,25], and two for the scam category [26,27]. Regarding spam, there were six studies [28–33], and five on fake profile detection [34–38]. Impersonation and other threat types were each addressed by one study. Next, we examined the techniques adopted by the selected works, including detection approaches, preprocessing procedures, and text representation methods.

The analysis of detection approaches revealed that text classification was the most prevalent method, employed in 14 studies. This aligns with the nature of the threats, as they primarily manifest in written communication. Four additional studies applied classification techniques to non-textual content. A few works introduced hybrid or alternative methods, including simulated annealing, a deep stacked autoencoder combined with DL, federated learning, a hybrid approach using rule-based and ML, sentiment analysis, topic modeling combined with classification, statistical matching, and a combination of classification and regression tasks.

Going into a little more detail, we sought to verify how the techniques were distributed among the most represented threat types. Studies that worked on phishing detection employed LLMs with prompts and contextual embeddings [16,17], ML with RF, SVM, etc. [19,20], and deep learning with autoencoders [18]. For Smishing-specific studies, techniques ranged from ML (RF with TF-IDF) [21] to DL (CNN-LSTM) [22] and hybrid TF-IDF/embedding models [23]. Social engineering threats were addressed using two approaches: LLM pipelines, using Retrieval-Augmented Generation (RAG) contextual embeddings [25], and ensemble DL models with bi-LSTM/CNN/MLP [24]. In spam detection, traditional ML methods remain highly prevalent. Approaches in this category included ML (NB, RF, SVM) [28,32,33], DL (LSTM, RNN) [31,32], ML+DL hybrid [29], and Federated Learning (with PhoBERT) [30]. The use of federated learning is particularly interesting as it can help to partially mitigate some privacy concerns. Fake profile detection employed diverse techniques, ranging from statistical methods [38] and traditional ML [36] to modern gradient-boosted ensembles [37], deep learning (RNN) [34], and Transformer-based models for tweet content analysis [35]. While text-based approaches focused on profile bios and tweets, others leveraged behavioral or network-level features.

Regarding preprocessing, it is noteworthy that ML-based studies provided more detailed documentation of these steps, which typically followed a consistent pipeline, including lowercasing, punctuation and stop-word removal, tokenization, and lemmatization. Some studies that report this treatment are [19,20,28,32,39,40]. Other mentioned techniques were emoji removal [23,41] or numerical feature scaling and categorical encoding for profile metadata [36,37].

Deep neural network-based approaches tended to employ fewer preprocessing transformations, or authors provided less detail about this stage in their methodology descriptions. LLMs typically bypassed manual preprocessing steps, instead relying on prompt engineering [16,17,26] or contextual embeddings [16,25].

An examination of preprocessing techniques by targeted threat type indicated that standard NLP preprocessing steps were widely employed in the majority of text message-based threat detection studies. For threats like fake profile or payment fraud, which often involve structured data or metadata, preprocessing steps shifted towards numerical/categorical conversion of attributes. While many steps were general, some were tailored:

- Slang removal in hacker discussions [42] indicated domain-specific cleaning;
- Length-based filtering for impersonation [41] suggested message structure was also important in this threat category.

Traditional lexical features were the most frequent text representation methods listed, aligning with the common use of feature-based ML approaches. This group of features included Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF),

n-grams, and message-level features (such as sentence length or structure, and uppercase usage), according to the second to last column of Table A1. Additionally, three works employed pre-trained word embeddings (GloVe and PhoBERT) [29,30,34], while contextual embeddings were mentioned seven times. Furthermore, several studies employed combined features, integrating lexical features with either embeddings [23,24,31] or semantic [27,43] and behavioral features [24,41].

Now, characterizing the text representation methods per threat category, we found that phishing and scam detection methods utilized both modern embeddings [16,17,26] and traditional techniques such as TF-IDF [19,20]. Static and contextual embeddings were used for text-based fake profile detection [34,35], in addition to other techniques with no text representation [36,37]. For the spam category, the textual representation ranged from simple BoW [28,32] to sophisticated embeddings [29,30]. Smishing studies employed both frequency-based representations [21] and SMS domain-adapted embeddings [22]. For social engineering, we observed a distinctive aspect: the use of a combination of modern representation techniques [24,25], including RAG. This may indicate that the authors aimed to capture not just superficial features of a message but also deeper semantic, contextual, and behavioral patterns.

4.3. Analysis of the Use of LLMs

As can be seen in the fourth column of Table A1, six out of thirty studies selected in this review, covering a period of one year and three months, employed LLM-based approaches. These publications specifically address threat categories phishing [16,17], scam [26], social engineering [25], harmful content [44], and one paper focused on hacker discussions as a early threat indicator [42]. Two additional studies employed a Transformer-based text classification approach for fake profile [35], and for social engineering [24] detection.

To illustrate the evaluation of two phishing detection approaches, studies [16,17] report accuracies of 100% and 97.5%, respectively. These results compare to 92% in [18] (DL-based) and 99% in [19] (ML-based) for the same threat type, though trained on distinct datasets.

LLMs demonstrate strong performance in context-intensive threats such as phishing and social engineering, primarily due to their advanced semantic understanding capabilities. However, LLMs are rarely used for behavior-centric threats (e.g., fake profile) or low-resource domains (e.g., Smishing).

Occasionally, the use of LLMs for generating simulated scenarios was noted as a limitation, as in [25], due to the risk of introducing unrealistic elements that could affect the reliability of the dataset.

Regarding language support in LLM-based studies, when specified, only English was mentioned. Multilingual capabilities were not reported in these six studies.

In half of the works involving LLMs, E-mail was the detection input scope, while the others took chat conversation history, posts, and comments as input.

Unexpectedly, among the 24 works where the methodology did not involve LLMs, only two instances [29,41] explicitly stated the intention to incorporate or investigate the application of LLMs within their future directions. Other studies indirectly alluded to more advanced models or had already employed LLMs and indicated future experiments with fine-tuned variants.

4.4. Analysis of Threat Detection Evaluation Methods

Out of 30 studies listed, 29 reported some form of evaluation, while 1 study explicitly stated "Planned, not performed" [40]. Hold-out validation and cross-validation were commonly employed. The fourth column (Evaluation) in Table A2 presents the results

reported by authors regarding their approaches' performance. The most frequently adopted metrics were accuracy, F1-score, precision and recall.

By combining those results with the Approach and Model Architecture (third and fourth) columns from Table A1, we can identify the peak reported performance for each detection technique, as presented in Table 2. Notably, within the phishing threat category, the top-performing study relies on traditional machine learning [19] rather than large language models.

It is also worth highlighting the assessment results by type of threat. In the case of fake profile detection, three studies reported an accuracy above 91% [34,36,38], while one study showed very limited performance [35]. Regarding phishing, the reported accuracy ranged from 62.5% to 100% in studies [16–20]. Three of the works on spam detection [28,30,32] reported an accuracy of 98% or higher, while two others presented values of 80.6% [29] and 85.6% [31]. The detection input scope for these first three was SMS, while for the others, it was described as textual messages.

Based on the reported evaluations, the studies generally demonstrated a high degree of effectiveness in the specific tasks they addressed, particularly in areas like SMS spam detection, phishing E-mail identification, and malicious chat content detection. However, it is important to note the considerable range in reported metrics. Studies like [17] (accuracy: 62.5% to 97.5%, F1: 44% to 97.6%) and [35] (precision: 31%, recall: 50%) highlight that effectiveness can vary greatly depending on the specific model, dataset quality, data size, and experimental setup.

The Limitations column in Table A2 provides additional context to put some results into perspective (e.g., "restricted data access" for [21], "small dataset" for [35], "dataset is focused on simulated scenarios" and "generated messages may be unrealistic" for [25]). These limitations suggest that real-world applicability or generalizability might be lower than the reported metrics imply in some cases.

Table 2. Peak performance by technique across all threat types.

Technique	Best Evaluation	Best-Case Threat Type	Top Citation
ML	Accuracy: 99.15%	Phishing (E-mail)	[19]
DL	Accuracy: 99.98%	Smishing (Swahili)	[22]
LLMs	Acc.: 97.5%, F1: 97.6%	Phishing	[17]
Ensemble DL	F1: 91.61%	Aggressive content	[43]
Statistical/other	Acc.: 91.6% to 100%	Fake profile	[38]

4.5. Analysis of Detection Input Scope

Considering the last column of Table A1, we note that the features considered by the models essentially characterized the message itself, and occasionally the message along with the preceding conversation history. Additional aspects, such as behavioral features and social network cues, were mentioned only in works addressing fake profile and impersonation detection.

Regarding the detection scope, that is, the category of data source being processed, as shown at the top of Figure 3, most studies focused on chats, forums, messaging platforms, SMS, and E-mail. A detailed analysis of the second column in Table A2 revealed the following: Nine instances of message history analysis (covering chats, forums, and unidirectional messaging platform channels); Eight instances of SMS message processing; Five cases of E-mail content examination; Two studies analyzing tweets (X platform); Two works incorporating profile account attributes; One instance of website content analysis; One study utilizing users' typing patterns.

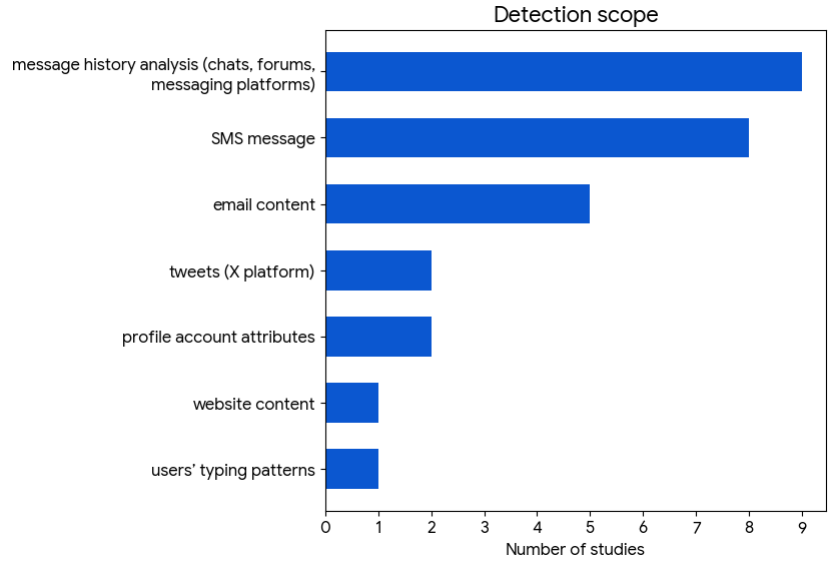

Figure 3. Distribution of data sources on which studies performed threat detection.

4.6. Analysis of Real-World Applicability and Ethical Compliance

Three of the selected studies [21,27,39] mentioned the possibility of threat detection in real time. The others were either silent on this aspect or explicitly indicated that it was not supported.

Among the 30 papers analyzed, 22 did not mention any ethical considerations related to real-world deployment, as shown in Table A3. These AI-based approaches involved message analysis and the automated processing of behavioral and social connection data. No explicit references to regulations such as GDPR were found.

An ethical dilemma is explicitly stated in [25]: publicly releasing the dataset could enable malicious use by certain actors.

In study [41], the authors report that personally identifiable information was removed to protect user privacy. The same work also documented that the detected fake and cloned channels were reported to Telegram.

In a few isolated cases, consent is mentioned, but only regarding direct participation in the study itself [38], not for potential future use of the system in real-world scenarios.

The work presented in [42] identifies privacy concerns and suggests corresponding risk mitigation strategies. The authors of [30] propose a federated learning approach to reduce sensitive information exposure. In [17], self-hosting a model is suggested for privacy, rather than using a cloud-deployed model and expanding the data perimeter.

5. Discussion

Following the synthesis of results, as outlined in the previous section and thoroughly detailed in Appendix A, we now present a series of considerations, addressing each research question individually.

RQ1: What NLP techniques are employed in threat detection for messaging platforms, and how are LLMs integrated into these approaches, according to recent research?

Threat detection for messaging platforms involves mainly classification tasks, typically binary, based on textual information. Sometimes, in addition to the text, other inputs are also considered, of a quantitative and structured nature, regarding properties of the profile of the message-sending interlocutor or their visible behavior. Graph-type data, related to connections on social networks, for example, may also exist. We also found distinct approaches for detection that involved regression, federated learning, hybrid solutions with ensembles, or approaches that combine heuristics with classification or with sentiment analysis.

LLMs are still not very prevalent in recent publications on the detection of threats related to messages and online interaction, with our review showing a weight of 6 out of 30, or a presence in 20% of the works. Reports on studies not using LLMs do not mention reasons for not adopting these tools. As many powerful LLMs reside in the cloud, and are controlled by private entities, sending sensitive data off-premises introduces significant risks, such as the lack of control on data retention policies, data interception risk, or the potential for the model to generate improper output, inadvertently exposing data related to the input provided [12,45,46]. This is a particular concern under stringent regulations like the GDPR. These data privacy challenges may perhaps explain part of the limited adoption of LLMs in the reviewed studies. The alternative of training a dedicated LLM, or using an on-premises open model, can mitigate privacy and legal compliance risks but demands significant computational resources. This can also be a discouraging factor for the use of LLMs in this context.

Focusing now on how LLMs are employed, models are used for the classification task central to detection, for generating clues and embeddings to be considered by another part of a deep neural network system, for the identification of topics or named entities or to support an auxiliary task, or for synthesizing new training data when original data are limited [25].

RQ2: What evaluation strategies are used to measure the effectiveness of these techniques?

The evaluations presented in the articles are scientifically sound, employing methods and metrics widely used in machine learning, particularly in the field of NLP. Since most detection approaches are classification-based, it is common practice to use hold-out validation or K-fold cross-validation on reference datasets. Performance is typically measured using accuracy, precision, recall, and F-measure.

In the case of regression-based approaches, the most commonly used evaluation metrics are the Mean Absolute Error (MAE—the average of the absolute differences between the predicted and actual values) and Mean Squared Error (MSE—the average of the squared differences between the predicted and actual values).

However, comparability across studies is challenging, as the datasets used often differ. In many cases, researchers need to create new datasets to address issues related to the sensitivity of existing data. As noted in one of the reviewed studies, this process of generating training instances can introduce bias into the dataset, potentially impacting the model's performance.

An important insight we can draw is that traditional machine learning remains highly competitive. For certain complex threats, such as phishing, ML approaches can even outperform LLMs, as shown in Table 2. Another notable aspect is the significant performance variability within threat categories. Even for the same threat type (e.g., fake profile or phishing), reported results vary considerably.

Globally, in this context of heterogeneous threats, effectiveness appears to be task-specific and input-dependent. There also seems to be a correlation between input scope (and consequently, task complexity) and the detection performance. To illustrate, high-performing spam detection studies primarily used SMS input, which may be simpler to process than longer conversation histories.

As seen in Section 4.4, some study limitations presented in Table A2 suggest that part of the reported results may not fully generalize to real-world settings, due to differences between experimental and live environments.

RQ3: What types of input are used for online threat detection in recent studies?

Analyzed threat detection systems make use of the following input sources: textual messages (E-mail, SMS, messaging platforms), isolated or within a conversation history; web content displayed on websites; profile attributes on social media; or behavioral data

quantified over time or related to social connections. These input scopes primarily focus on various forms of text-based communication, and crucial non-textual data related to user profile and behavior. This diversification of input types reflects the evolving nature of cyberthreats and the need for more comprehensive detection systems that can analyze a wider array of digital footprints.

Perhaps it would be useful for the effectiveness of detecting certain types of threats, such as those associated with fake profiles, to consider a broader detection input scope than just the content of the message or a one-sided view. Studies in social engineering detection often demand a deep understanding of the context, particularly the conversation history. On the other hand, the collaboration between human analysts and automatic systems, and therefore a new source of input for multimodal AI systems, can contribute to the detection performance.

RQ4: How can detection efficacy be balanced with privacy, informed consent, and compliance with legal and ethical regulations?

The findings of this review indicate that ethical implications related to the later deployment phase are not currently a priority during the research stage focused on designing and optimizing techniques, or at least, there is little to no documented discussion of this topic.

Concern for privacy and ethics in the use of AI is explicitly expressed in some of the analyzed works, with informed consent from participants being used in some studies. Likewise, the validation of the work referenced in certain papers by Institutional Review Board is pertinent; however, it does not obviate the necessity of devising a security, privacy, and ethics strategy for the application of AI upon real-world deployment.

Considering the implementation tools mentioned, the studies using ML and DL appear to minimize the scope of data exposure, whereas the approaches based on LLMs generally rely on remote services and platforms, rather than using open models running locally. Data transfers for remote processing create significant privacy and compliance obligations, particularly given the jurisdictional uncertainties inherent in deperimeterized environments.

In our rapidly evolving digital landscape, advancing digital literacy is crucial. This must encompass not only awareness of online threats but also an understanding of ethical implications and legal compliance requirements. An effective real-world solution must incorporate data subject rights from inception, respecting their authorizations or refusals regarding data use for each specific purpose and time period.

Privacy by design is recommended to minimize risks. Data protection is paramount, driven by cryptographic methods and robust access control. On-premises data processing minimizes data exposure. Where feasible, privacy-enhancing technologies like federated learning or homomorphic encryption should be implemented as complementary safeguards. Anonymization must be systematically applied, being particularly critical in the event of a system breach, serving to mitigate the risk to data subjects. In services involving data transfers, all flows must be mapped and compliance with the data sovereignty principles verified. Data processing agreements must be prepared when a service involves a third-party entity.

In the European Union, the Artificial Intelligence Act (AI Act (https://artificialintelligenceact.eu/ (accessed on 15 April 2025))) is a comprehensive regulatory framework for AI to ensure safety, transparency, and fundamental rights protection. An online compliance checker for new AI-related obligations is available. It also emphasizes the importance of AI literacy, listing some training programs that may be useful to better understand AI technologies and their responsible use.

6. Conclusions

The current threat detection literature largely lacks a transversal perspective on common message-based patterns. Furthermore, the practical applicability of these detection strategies for real-world citizen protection, especially concerning user privacy and regulatory compliance, requires deeper investigation. In response to these gaps, we designed this systematic review of NLP techniques for detecting message-based threats on digital platforms, conducting a bibliographical survey of recently published studies while following PRISMA guidelines. The review search and screening pipeline resulted in a set of 30 relevant articles, whose information was analyzed according to a defined protocol and recorded as shown in Appendix A.

Our main contributions include a survey of the strategies used to detect message-based threats and the role of large language models (LLMs), the evaluation methods and performance of the studied approaches, the types of data sources used for threat detection, and discussions surrounding privacy and ethics. To the best of our knowledge, this is the first study to quantitatively assess the prevalence of LLMs in recent research on message-based threat detection.

One limitation of our work may be the short time frame used for selecting publications. On one hand, a year and a quarter might seem like a short period. On the other hand, it seems unlikely that relevant papers published before 2024 would use open LLM models for threat detection. Additionally, extending the time frame backwards would reduce the up-to-date nature that was intended for this systematic review.

Regarding the use of LLMs, there still seems to be a way to include this type of model in message-based threat detection techniques. Perhaps the computational complexity, added to the ethical challenges in using message and interlocutor data, is the justification. This will be something to validate in a future extension of this work. The trend towards open LLM models, offering personalized fine-tuning and less centralized deployment, may soon facilitate a wider adoption of dedicated LLMs, thus avoiding reliance on third-party entities and deperimetrization issues.

For attackers, it is easier to try and find a victim by exploiting the mass sending of messages and resorting to automated systems. For defensive systems, it is difficult to guarantee that no point of vulnerability exists. And unlike attackers, the training of models and the use of data must follow good ethical and legal practices, which may add some delay to the progress of protection techniques.

When it comes to ethics in AI, the balance between data privacy and the ability to develop models for threat detection and timely protection is a challenge. However, while safeguarding the consent of the data subject, the important thing is to ensure that the impact of data exposure to AI is proportionate to the severity of the threat it aims to mitigate [47].

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUC	Area Under the ROC Curve

BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CTI	Cyberthreat Intelligence
DL	Deep Learning
DT	Decision Tree
F1	F1-Score (harmonic mean of precision and recall)
GPT	Generative Pre-trained Transformer
IRB	Institutional Review Board
KNN	K-Nearest Neighbors
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
PII	Personally Identifiable Information
RAG	Retrieval-Augmented Generation
RF	Random Forest
RNN	Recurrent Neural Network
SMS	Short Message Service (text message)
SVM	Support Vector Machine

Appendix A

This Appendix presents the tables containing data items collected during our analysis of the selected papers in this review, following the protocol outlined in Section 3.2.

Table A1. Summary of approaches and NLP techniques used in the selected studies.

Ref	Targeted Threats	Approach	Model Architecture	Preprocessing	Text Representation	Detection Features
[16]	Phishing	Text classification using LLMs	LLM	Prompt development	LLM's contextual embeddings	Text-based features: message content, subject, sender, recipient
[26]	Scam	LLM applies a scam detection rubric	LLM	Prompt development including instructions, E-mail, and rubric	LLM's contextual embeddings	Text-based features and rubric rules
[21]	Smishing	Text classification	ML (Random Forest)	Tokenization, stopword removal, stemming	TF-IDF	Text-based features
[39]	Toxic communication	Text classification	ML (RF, SVM, LR, MNB, KNN), and BERT	Normalization, tag and link removal, stopword removal, stemming, lemmatization	TF-IDF	Text-based features
[34]	Fake profile	Classification	DL (RNN: GRU+LSTM)	Tokenization, normalization	Word embedding: GloVe	Text-based features (from tweets, user profile, network)
[40]	Inappropriate messages	Rule-based, sentiment analysis	DL (LSTM)	Tokenization, stopword removal, stemming	Unspecified	Text-based features, behavior temporal features

Table A1. *Cont.*

Ref	Targeted Threats	Approach	Model Architecture	Preprocessing	Text Representation	Detection Features
[48]	Authenticity classification	Topic modeling (LDA), text classification	XLNet and BERT	Tokenization, special characters and stopword removal	TF-IDF with LDA; contextual embeddings for classification	Text-based features
[17]	Phishing	Benchmarking 12 LLMs instructed for detection	LLM	Benchmarking setup; prompt development	LLM's contextual embeddings	Text-based features
[35]	Fake profile	text classification	Transformer-based (BERT)	tag, punctuation and stopword removal; lemmatization	BERT's contextual embeddings	Text-based features from tweet content and properties
[24]	Social engineering	Simulated annealing optimized fusion of five specialized DL models	DL (bi-LSTM, CNN, MLP) and Transformer-based	Not mentioned. Specific to inner models.	Static embeddings and BERT's contextual embeddings	personality traits, linguistic aspects, behavioral characteristics, and IT attributes
[28]	Spam	text classification	ML (NB, RF, SVM…)	Tokenization, stemming	BoW and TF-IDF	Text-based features
[25]	Social engineering	LLM-based text classification pipeline	LLM	Not mentioned	Contextual embeddings (LLM+RAG)	Text-based features (message-level and conversation-level)
[29]	Spam	Text classification	ML, DL	Tokenization, stopword removal, stemming	GloVe embeddings	Text-based features
[18]	Phishing	DL+deep stacked autoencoder	DL	Not mentioned	BoW, n-gram, hashtags, sentence length, uppercase, TF-IDF	Web and URL features and text-based features
[30]	Spam	Federated learning for classification	FedAvg, FedAvgM, FedAdam	NA	PhoBERT's embeddings	Text-based features
[44]	Harmful content	LLMs as content classifiers in a custom policy moderation pipeline	LLM	NA	GPT 3.5 and LLaMa 2's embeddings	Text-based features
[36]	Fake profile	Classification	ML (Random Forest, SVM, KNN)	Converting profile attributes into numerical form	NA (the proposal uses no text)	Profile account metadata, behavioral features, network follower, and following numbers
[31]	Spam	text classification	DL (LSTM)	Punctuation, stopwords, and URL removal; lemmatization	BoW, TF- IDF, word embeddings	Text-based features
[37]	Fake profile, payment fraud	Classification	ML (XGBoost, CatBoost, GBM)	Numerical feature scaling, categorical encoding, derived features' generation.	Not mentioned	Average likes per post, sentiment analysis of profile bio text, post hashtags
[41]	Impersonation, clone channels	Classification	ML (RF, SVM) and DL (MLP)	Removal of mentions, numbers, links, emoji, and messages shorter than 15 characters; tokenization	Message format and structure features (not the usual text representation)	Behavioral features (#forwarded messages, average length of posted messages, #messages posted in the last 3 months, interaction features); profile attributes

Table A1. *Cont.*

Ref	Targeted Threats	Approach	Model Architecture	Preprocessing	Text Representation	Detection Features
[43]	Aggressive content	Text classification	ML, DL (LSTM)	Punctuation and stopword removal; stemming; Labeling	Semantic features (actor, target, polarity), and TF-IDF	Text-based features
[38]	Fake profile	Statistical methods for data match score	Separate and combined similarity verifiers	Not mentioned	NA	Key hold time, key interval time, word hold time
[22]	Smishing	Text classification	DL (CNN + LSTM)	Lowercase conversion, tokenization, punctuation and stopword removal	Trained embeddings	Text-based features
[27]	Malicious (spam, phishing, or fraud) messages	Text classification	ML	Stopword removal, stemming, lemmatization	Lexical (word frequencies, n-grams), Semantic features (fraud related), context features (sender, time related)	Text-based features and behavioral features
[42]	Hacker discussions as early threat indicator	LLM-based named entity recognition and classification	NLP, ML, and LLM	lowercase conversion; tokenization; punctuation, stopword, slang and non-ASCII removal; lemmatization	BoW and TF-IDF for SVM; BERT's contextual embeddings	Text-based features
[32]	Spam	Text classification	ML (LR, SVM, RF), DL (RNN)	Tokenization, stopword and punctuation removal, stemming	BoW, message length	Text-based features
[33]	Spam	Classification and regression approach	ML (RF, RF Regressor)	Not mentioned	Features representing the outcome of criteria checks (binary/numerical)	56 features extracted using NLP and grouped into categories: headers, text, attachments, URLs, and protocols
[20]	Phishing	Text classification	ML (RF)	Stemming via AraBERT	TF-IDF	Text-based features
[19]	Phishing	Text classification	ML (RF, DT, LR, SVM)	Convert to lowercase, punctuation and stopword removal, tokenization, lemmatization	TF-IDF	Text-based features
[23]	Smishing	Text classification	ML (LR, SVM, RF…)	Prior feature extraction, tokenization, emojis and punctuation removal, lowercase	TF-IDF and BERT embeddings	Metadata and text-based features

Table A2. Language, inputs, evaluation, research limitations and future directions reported in the selected studies.

Ref	Supported Language	Input Scope	Evaluation	Limitations	Future Directions
[16]	Unspecified (multilingual ability by LLM)	E-mail	Accuracy, confidence score distribution; OpenAI models excel	Sporadic experience involving copying text into a prompt for the analyzed models	Further model fine-tuning and refinement
[26]	Unspecified (multilingual ability by LLM)	E-mail	Accuracy: 69% to 98%	Limited experience; vulnerable to OpenAI interface downtime, and API rate limits	Refining the scam detection rubric; creating an entirely self-contained ML algorithm
[21]	English and Bemba	SMS text messages	F1: 90.2%, AUC: 95%	Restricted data access	The need for further model optimization
[39]	English	Chat message content	Accuracy: 84% to 92%	Not specified; scale challenges	Optimizing ML parameters, bot enhancing, and multilingual support

Table A2. *Cont.*

Ref	Supported Language	Input Scope	Evaluation	Limitations	Future Directions
[34]	Mostly English (dataset-dependent)	Tweets	Accuracy: 99.73%, Precision: 98.23%, Recall: 99.56%, F1: 99.63%	Not specified	Wider range of features, multilingual, language-agnostic features
[40]	Unspecified	Chat message content	Planned, not performed	Not specified; the proposed methodology lacks implementation details	Refine the core components; broader chat platform scope
[48]	Chinese	Chat messages' content	Accuracy: 77% to 82.5%	Not specified.	Optimizing the model and enhancing generalization, alongside supporting diverse languages and cultural backgrounds
[17]	English	E-mail	Accuracy: 62.5% to 97.5%, F1: 44% to 97.6%	The benchmark used base models and small datasets; optimized temperature value was different for each model	Phishing detection using fine-tuned LLMs; expanding datasets to include more complex scenarios
[35]	English (multilingual depends on BERT version)	Tweets	Precision: 31%, recall: 50%	Not specified; poor performance; small dataset	Not specified
[24]	English	Chat history	Accuracy: 79.9%, AUC: 74.3%, F1: 70.1%	Gap between the AUC value of the multimodal fusion model and one inner component; further research is needed to investigate the model interpretability	Expand capabilities, identifying deepfake content
[28]	English	SMS text messages	Accuracy: 98%	Scalability challenges; dataset size	Exploring temporal patterns features
[25]	English; multilingual not specified	Chat message; chat conversation (history)	F1: 80%	Dataset is focused on simulated scenarios in a particular topic; LLM-generated messages may be unrealistic or overly agreeable and could affect the dataset reliability	Expand to other domains such as financial services or customer support; consider the broader ethical and practical LLM usage implications
[29]	English	Message text	Accuracy: 80.6%, F1: 75.5%	Not specified	Incorporate BERT or GPTs; use HuggingFace
[18]	Unspecified	Website data	Accuracy: 92%, TPR: 92.5%, TNR: 92.1%	Not specified	Utilize transfer learning
[30]	Vietnamese	SMS text messages	Accuracy: 98%	NA	NA
[44]	English	Posts, comments, messages (text)	F1: 44% to 77.6%	Cannot differentiate between a violation detected with high confidence and one just suspected	Expand across diverse languages, cultures, and contexts, even involving real users; using multimodal models, generalize the approach to different types of data
[36]	Unspecified	Profile account attributes	Accuracy: 92.5%	Not specified	Not specified
[31]	Unspecified (dataset dependent)	Message	Accuracy: 85.6%	Not specified	Contextual analysis; user-centric approaches.
[37]	Unspecified	Profile account attributes	Accuracy: 93%, F1: 92%	Not specified	Detection across various domain sectors
[41]	English	Channel text messages; message-related counters; channel (profile) attributes	F1: 85.45%	Focused only on English channels	Incorporate LLMs and include semantic features.

Table A2. *Cont.*

Ref	Supported Language	Input Scope	Evaluation	Limitations	Future Directions
[43]	English	Text sentences	F1: 77.13% for LR/ML, F1: 91.61% for BiLSTM/DL	Linguistic features seem to make the models over-generalize; for the textual features, models tend to be biased; model interpretability; limited to English.	Dataset's human annotation; increase dataset positive instances
[38]	Unspecified	Users' typing patterns	Accuracy: 91.6% to 100%	Not specified	Augment the dataset with additional users; investigate additional linguistic features; include multimodal, multi-device, usage-context, and DL
[22]	Swahili	SMS text messages	Accuracy: 99.98%	Not specified	Tune hyperparameters to reduce false negatives; develop a mobile application
[27]	English	SMS, chat, app, or E-mail messages	Accuracy: 94.6%, F1: 93.2%	Not specified	Not specified
[42]	Dataset-dependent (English)	Forum or group chat contents	Accuracy: 91.7%, F1: 87.8%	Communities are probably moving to other platforms; sampling bias; report and forum information reliability; the scope of the analysis was English-speaking forums only	Develop benchmark datasets for NER models within the CTI domain; focus on topic matching that incorporates contextual semantics; extend data sources to include modern platforms
[32]	English	SMS text messages	Accuracy: 99.28%	Not specified	Train the model with other datasets and fine tuning; curate new SMS datasets
[33]	English and Spanish	E-mail message with headers	F1: 91.4% for classification, and MSE: 0.781 for regression	Considerable feature extraction time	Development of an ensemble model based on stacked generalization; explore more distinctive features
[20]	Arabic	SMS text messages	Accuracy: 98.66%, F1: 98.67%	Some inaccuracies in translating an English dataset to Arabic	Use a real Arabic dataset
[19]	Dataset-dependent (English)	E-mail	Accuracy: 98.72% and 99.15%, depending on the dataset	Not specified	Dataset augmentation; try a wider range of phishing strategies and linguistic nuances
[23]	Unspecified	SMS text messages	Accuracy: 94%, F1: 93.78%	Not specified	Extend to analyze various social platform messages; collect other language-based datasets.

Table A3. Platform, datasets, tools, and ethical aspects reported in the reviewed studies.

Ref	Deployment Platform	Datasets	Real-Time	Implementation Tools	Ethical Aspects of Real-World Deployment
[16]	E-mail	Fraudulent E-mail corpus	N	poe.com AI chat platform, LLMs	Not mentioned
[26]	E-mail	Nazario database, Untroubled Scam 2023 Archive, and custom	N	OpenAI ChatGPT API, ChatGPT 3.5	Not mentioned
[21]	SMS	English + Bemba Smishing datasets	Y	Unspecified	Not mentioned
[39]	Telegram platform	Chat dataset from social platforms and Kaggle	Y	HuggingFace, Python packages	Not mentioned
[34]	X platform (Twitter)	TwiBot-20 dataset	Unspecified	Python packages	Not mentioned
[40]	Messaging apps	Unspecified	NA	Unspecified	Data anonymization and user consent are proposed
[48]	WeChat group chats	Custom chat dataset	Unspecified	HuggingFace's Transformers library	Not mentioned

Table A3. *Cont.*

Ref	Deployment Platform	Datasets	Real-Time	Implementation Tools	Ethical Aspects of Real-World Deployment
[17]	E-mail	Enterprise E-mails (contemporary message quality)	NA	Chatbot Arena, Ollama	Self-hosting a model is suggested for privacy
[35]	X platform (Twitter)	Collected tweet dataset	Unspecified	Unspecified	Not mentioned
[24]	Chat platforms	Chat social engineering corpus	Unspecified	SpaCy, PyTorch, HuggingFace	Not mentioned
[28]	SMS	SMS Spam Collection	N	Python sklearn	Not mentioned
[25]	Chat platforms	SEConvo, a developed dataset	Unspecified	LangChain, OpenAI API, Faiss, Python	Ethical dilemma: the potential dataset misuse; privacy and consent are not mentioned.
[29]	SMS, chat, messaging platforms	Compilation of SMS and Twitter datasets	N	Unspecified	Not mentioned
[18]	Content referring Webpages	Webpage phishing detection dataset	N	Python	Not mentioned
[30]	SMS	Spam dataset	Unspecified	NA	FL is used to reduce the exposure of sensitive information
[44]	Social media	OpenAI's content moderation dataset; Reddit's Multilingual Content Moderation dataset	Unspecified	Unspecified	Ethical and legal concerns are reported on how data and personal information might be used for training without the user's awareness
[36]	Instagram platform	Instagram fake/spammer/genuine accounts	Unspecified	Python, Pandas, NumPy, Scikit-learn	Not mentioned
[31]	Social media	Custom dataset	Unspecified	NLTK, Pandas, TensorFlow, PyTorch, Scikit-learn	Not mentioned
[37]	Social Media	Fake profile and payment fraud datasets	Unspecified	Unspecified	Not mentioned
[41]	Telegram platform	Collected data from 120,979 Telegram public channels	Unspecified	NLTK, LangDetect, Scikit-learn, PyTorch, Telegram API	The protection of user privacy is mentioned; PII was removed; authors reported the detected fake and clone channels to Telegram
[43]	Text message platform	LLM generated dataset	Unspecified	Textblob, ChatGPT, Google Colab	Not mentioned
[38]	Social networks	Collected keystroke timings on Facebook, X, and Instagram	Unspecified	Python	Study participants signed a consent; deployment phase not mentioned
[22]	SMS	32,259 Swahili SMS messages	Unspecified	Python, Keras, Google Colab	Not mentioned
[27]	Mobile messaging applications	50,000 (benign, spam, phishing, and fraud) messages	Y	Unspecified	Not mentioned
[42]	Online forums	Hacker forums articles covering 20 years; threat report data (intelligence community, press, sites)	N	NLTK, DarkBERT	Ethical and privacy concerns are mentioned by the authors regarding the use of conversations; some measures are suggested to mitigate privacy risks
[32]	SMS	SMS Spam Dataset (UCI)	Unspecified	Python, Google Colab	Not mentioned
[33]	E-mail	Custom dataset built from two sources	N	Python, Scikit-learn	Not mentioned
[20]	SMS	Custom, translated dataset: 638 phishing and 4,844 legitimate messages	Unspecified	AraBERT (Arabic LM), Python, Scikit-learn, Google Colab	Not mentioned
[19]	E-mail	Fraud E-mail dataset and phishing E-mail dataset	Unspecified	Unspecified	Not mentioned
[23]	SMS	Custom dataset combining messages from three sources	Unspecified	NLTK, Scikit-learn, TensorFlow	Not mentioned

References

1. Thumboo, S.; Mukherjee, S. Digital romance fraud targeting unmarried women. *Discov. Glob. Soc.* **2024**, *2*, 105. [CrossRef]
2. Buil-Gil, D.; Zeng, Y. Meeting you was a fake: Investigating the increase in romance fraud during COVID-19. *J. Financ. Crime* **2022**, *29*, 460–475. [CrossRef]
3. Alharbi, A.; Dong, H.; Yi, X.; Tari, Z.; Khalil, I. Social Media Identity Deception Detection: A Survey. *Acm Comput. Surv.* **2021**, *54*, 69. [CrossRef]
4. Perik, L.W. Leveraging Generative Pre-trained Transformers for the Detection and Generation of Social Engineering Attacks: A Case Study on YouTube Collusion Scams. Master's Thesis, University of Twente, Enschede, The Netherlands, 12 January 2025.
5. Kyaw, P.H.; Gutierrez, J.; Ghobakhlou, A. A Systematic Review of Deep Learning Techniques for Phishing Email Detection. *Electronics* **2024**, *13*, 3823. [CrossRef]
6. Europol. *Internet Organised Crime Threat Assessment (IOCTA) 2021*; Publications Office of the European Union: Luxembourg , 2021. [CrossRef]
7. Europol. *Internet Organised Crime Threat Assessment (IOCTA) 2024*; Publications Office of the European Union: Luxembourg, 2024. [CrossRef]
8. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 3rd ed.; Stanford University: Stanford, CA, USA, 2025.
9. IBM. What is Machine Learning? Available online: https://www.ibm.com/think/topics/machine-learning (accessed on 4 April 2025).
10. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
12. Barberá, I. AI Privacy Risks & Mitigations—Large Language Models (LLMs). European Data Protection Board. Available online: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf (accessed on 12 June 2025).
13. Booth, A.; Sutton, A.; Clowes, M.; James, M. *Systematic Approaches to a Successful Literature Review*, 3rd ed.; SAGE Publications Ltd.: Thousand Oaks, CA, USA, 2022.
14. Fakhouri, H.; Alhadidi, B.; Omar, K.; Makhadmeh, S.; Hamad, F.; Halalsheh, N. AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response. In Proceedings of the 2nd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 26–28 February 2024; pp. 1–8. [CrossRef]
15. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst. Rev.* **2022**, *18*, e1230. [CrossRef] [PubMed]
16. Patel, H.; Rehman, U.; Iqbal, F. Evaluating the Efficacy of Large Language Models in Identifying Phishing Attempts. In Proceedings of the 16th International Conference on Human System Interaction (HSI), Paris, France, 8–11 July 2024; pp. 1–7. [CrossRef]
17. Zhang, J.; Wu, P.; London, J.; Tenney, D. Benchmarking and Evaluating Large Language Models in Phishing Detection for Small and Midsize Enterprises: A Comprehensive Analysis. *IEEE Access* **2025**, *13*, 28335–28352. [CrossRef]
18. Vidyasri, P.; Suresh, S. FDN-SA: Fuzzy deep neural-stacked autoencoder-based phishing attack detection in social engineering. *Comput. Secur.* **2025**, *148*, 104188. [CrossRef]
19. Aljamal, M.; Alquran, R.; Aljaidi, M.; Aljamal, O.S.; Alsarhan, A.; Al-Aiash, I.; Samara, G.; Banisalman, M.; Khouj, M. Harnessing ML and NLP for Enhanced Cybersecurity: A Comprehensive Approach for Phishing Email Detection. In Proceedings of the 25th International Arab Conference on Information Technology, Zarqa, Jordan, 10–12 December 2024. [CrossRef]
20. Ibrahim, A.; Alyousef, S.; Alajmi, H.; Aldossari, R.; Masmoudi, F. Phishing Detection in Arabic SMS Messages using Natural Language Processing. In Proceedings of the Seventh International Women in Data Science Conference at Prince Sultan University, Riyadh, Saudi Arabia, 3–4 March 2024. [CrossRef]
21. Zimba, A.; Phiri, K.; Kashale, C.; Phiri, M. A machine learning and natural language processing-based smishing detection model for mobile money transactions. *Int. J. Inf. Technol. Secur.* **2024**, *16*, 69–80. [CrossRef]
22. Mambina, I.S.; Ndibwile, J.D.; Uwimpuhwe, D.; Michael, K.F. Uncovering SMS Spam in Swahili Text Using Deep Learning Approaches. *IEEE Access* **2024**, *12*, 25164–25175. [CrossRef]
23. Jain, A.K.; Kaur, K.; Gupta, N.K.; Khare, A. Detecting Smishing Messages Using BERT and Advanced NLP Techniques. *SN Comput. Sci.* **2025**, *6*, 109. [CrossRef]
24. Tsinganos, N.; Fouliras, P.; Mavridis, I.; Gritzalis, D. CSE-ARS: Deep Learning-Based Late Fusion of Multimodal Information for Chat-Based Social Engineering Attack Recognition. *IEEE Access* **2024**, *12*, 16072–16088. [CrossRef]
25. Ai, L.; Kumarage, T.; Bhattacharjee, A.; Liu, Z.; Hui, Z.; Davinroy, M.; Cook, J.; Cassani, L.; Trapeznikov, K.; Kirchner, M.; et al. Defending Against Social Engineering Attacks in the Age of LLMs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–26 November 2024; pp. 12880–12902. [CrossRef]

26. DiMario, C.L.; Bacha, R.C.; Butka, B.K. Combatting Senior Scams Using a Large Language Model-Created Rubric. In Proceedings of the 5th Asia Service Sciences and Software Engineering Conference (ASSE 2024), Tokyo, Japan, 11–13 September 2024. [CrossRef]
27. Reddy, M.; Pallerla, R. Using AI to Detect and Classify Suspicious Mobile Messages in Real Time. In Proceedings of the 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 5–7 February 2025; pp. 1772–1777. [CrossRef]
28. Dharrao, D.; Gaikwad, P.; Gawai, S.V.; Bongale, A.M.; Patel, K.; Singh, A. Classifying SMS as spam or ham: Leveraging NLP and machine learning techniques. *Int. J. Saf. Secur. Eng.* **2024**, *14*, 289–296. [CrossRef]
29. Asmitha, M.; Kavitha, C.R. Exploration of Automatic Spam/Ham Message Classifier Using NLP. In Proceedings of the 9th International Conference for Convergence in Technology (I2CT), Pune, India, 5–7 April 2024; pp. 1–7. [CrossRef]
30. Anh, H.Q.; Anh, P.T.; Nguyen, P.S.; Hung, P.D. Federated Learning for Vietnamese SMS Spam Detection Using Pre-trained PhoBERT. In Proceedings of the 25th International Conference on Intelligent Data Engineering and Automated Learning—IDEAL 2024, Valencia, Spain, 20–22 November 2024; Volume 15346. [CrossRef]
31. Sivakumar, M.; Abishek, S.A.; Karthik, N.; Vanitha, J. Offensive Message Spam Detection in Social Media Using Long Short-Term Memory. In Proceedings of the 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), New Delhi, India, 21–23 November 2024. [CrossRef]
32. Bennet, D.T.; Bennet, P.S.; Thiagarajan, P.; Sundarakantham, K. Content Based Classification of Short Messages using Recurrent Neural Networks in NLP. In Proceedings of the International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Victoria, Seychelles, 1–2 February 2024; pp. 1–6. [CrossRef]
33. Jáñez-Martino, F.; Alaiz-Rodríguez, R.; González-Castro, V.; Fidalgo, E.; Alegre, E. Spam email classification based on cybersecurity potential risk using natural language processing. *Knowl.-Based Syst.* **2025**, *310*, 112939. [CrossRef]
34. Geetha, B.; Sushmitha, B.; Ilanchezhian, P.; Alabdeli, H.; Ahila, R. A Bi-directional Gated Recurrent Unit and Long Short-Term Memory based Fake Profile Identification System. In Proceedings of the First International Conference on Software, Systems and Information Technology (SSITCON), Tumkur, India, 18–19 October 2024; pp. 1–5. [CrossRef]
35. Singha, A.K.; Paul, A.; Sonti, S.; Guntur, K.; Chiranjeevi, M.; Dhuli, S. BERT-Based Detection of Fake Twitter Profiles: A Case Study on the Israel-Palestine War. In Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–6. [CrossRef]
36. Arunprakaash, R.R.; Nathiya, R. Leveraging Machine Learning algorithms for Fake Profile Detection on Instagram. In Proceedings of the 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 8–9 August 2024; pp. 869–876. [CrossRef]
37. Asha, V.; Nithya, B.; Prasad, A.; Kumari, M.; Hujaifa, M.; Sharma, A. Optimizing Fraud Detection with XGBoost and CatBoost for Social Media Profiles and Payment Systems. In Proceedings of the International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 11–13 February 2025; pp. 1987–1992. [CrossRef]
38. Kuruvilla, A.; Daley, R.; Kumar, R. Spotting Fake Profiles in Social Networks via Keystroke Dynamics. In Proceedings of the IEEE 21st Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 6–9 January 2024; pp. 525–533. [CrossRef]
39. Abhijith, A.B.; Prithvi, P. Automated Toxic Chat Synthesis, Reporting and Removing the Chat in Telegram Social Media Using Natural Language Processing Techniques. In Proceedings of the Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 11–12 January 2024; pp. 1–7. [CrossRef]
40. Shiny, J.; Penyameen, S.; Hannah, N.; Harilakshmi, J.S.; Hewin, A.; Thanusha, S. Analysis of Behavior in Chat Applications using Natural Language Processing. In Proceedings of the 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 10–12 July 2024; pp. 718–725. [CrossRef]
41. La Morgia, M.; Mei, A.; Mongardini, A.; Wu, J. Pretending to be a VIP! Characterization and Detection of Fake and Clone Channels on Telegram. *Acm Trans.* **2024**, *19*, 1–24. [CrossRef]
42. Paladini, T.; Ferro, L.; Polino, M.; Zanero, S.; Carminati, M. You Might Have Known It Earlier: Analyzing the Role of Underground Forums in Threat Intelligence. In Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '24), Padua, Italy, 30 September–2 October 2024; pp. 368–383. [CrossRef]
43. Raza, M.O.; Meghji, A.F.; Mahoto, N.A.; Reshan, M.S.A.; Abosaq, H.A.; Sulaiman, A.; Shaikh, A. Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 183. [CrossRef]
44. Franco, M.; Gaggi, O.; Palazzi, C.E. Integrating Content Moderation Systems with Large Language Models. *Acm Trans.* **2024**, *19*, 1–21. [CrossRef]
45. Feretzakis, G.; Vagena, E.; Kalodanis, K.; Peristera, P.; Kalles, D.; Anastasiou, A. GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet* **2025**, *17*, 151. [CrossRef]

46. Narayan, S.M.; Kohli, N.; Martin, M.M. Addressing contemporary threats in anonymised healthcare data using privacy engineering. *NPJ Digit. Med.* **2025**, *8*, 145. [CrossRef] [PubMed]
47. Karliuk, M. Proportionality principle for the ethics of artificial intelligence. *AI Ethics* **2023**, *3*, 985–990. [CrossRef]
48. Nie, N.; Guo, H.; Song, W. Authenticity Classification of WeChat Group Chat Messages Based on LDA and NLP. In Proceedings of the 9th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 25–27 April 2024; pp. 313–319. [CrossRef]